A nucleic acid encoding a geranyl-geranyl pyrophosphate synthetase (GGPPS) and polymorphic markers associated with said nucleic acid.

FIELD OF THE INVENTION

The present invention relates to a purified or isolated polynucleotide encoding human geranylgeranyl pyrophosphate synthetase, the regulatory nucleic acids contained therein, a polymorphic marker thereof and the resulting encoded protein, as well as to methods and kits for detecting this polynucleotide and this protein. The present invention also pertains to a polynucleotide carrying the natural regulatory regions of the *hGGPS* gene which is useful, for example, to express a heterologous nucleic acid in host cells or host organisms as well as functionally active regulatory polynucleotides derived from said regulatory region. The invention also consists in genetic markers, namely biallelic markers, which may be useful for the diagnosis of diseases related to an alteration in the regulatory or coding regions of *hGGPS*, such as pathologies related to a defect in the mevalonic biosynthetic pathway.

BACKGROUND OF THE INVENTION

Prenylation is the least common known lipid modification. Other lipid modifications include palmitylation, myristylation and glycophospholipidation. However, prenylation is a surprisingly common form of post-translational protein modification with an occurrence of 0.5 % of all cellular proteins. Prenylation is a covalent modification which involves the attachment of either a C15 farnesyl or a C20 geranylgeranyl isoprenoid, both being products of the mevalonic acid biosynthetic pathway, to one or more cysteine residues at the carboxyl terminus of the protein via a thioether bond. The C20 geranylgeranyl modification predominates over the C15 farnesyl modification in terms of frequency of occurrence. The structural environment of the cysteine residue determines the specific type and number of isoprenoid groups that attach to each cysteine. The covalent modification resulting from prenylation renders proteins more hydrophobic and, together with a subsequent modification cascade, facilitates their association with membranes. Protein prenylation also mediates protein-protein interactions. Prenylated proteins can be involved in signal transduction, intracellular vesicular transport, cytoskeletal organization, cell growth control and polarity, viral replication and protein folding/assembly. In mammals, prenylated proteins are more frequently modified by one or more geranylgeranyl groups. Farnesylation has only been found to occur in the retinal heterotrimeric G protein transducin, in retinal rhodopsin kinase, in *ras* proteins, in nuclear lamins, and in yeast mating factors. Geranylgeranylation is found in all of the remaining heterotrimeric G proteins and small G proteins.

Heterotrimeric G-proteins which are required for intracellular signal transduction between receptors and effector enzymes present one or two prenylated subunits. This modification is often required for association of the functional complex with the membrane.

Among small G proteins, *Ras* proteins, which comprise oncogenic forms, regulate signal transduction pathways controlling cell proliferation and differentiation. All *ras* proteins are prenylated and this modification is critical for their transport to the inner surface of the plasma membrane and their biological functions.

Other prenylated proteins belonging to the *ras* protein superfamily are involved in the regulation of intracellular vesicular transport (Rab/YPT1), in the cytoskeletal organization of polymerized actin to produce stress fibers (Rho) or membrane ruffling (Rac), in the oxydative burst of phagocytic cells (Rac), in the control of the cell cycle and polarity (cdc24Hs/G25K), and in negative growth control (Rap/Krev-1). Prenylation is important to these activities. For example, Rab/YPT prenylation is critical for the association of these proteins with specific intracellular compartments and in their regulation of intracellular transport processes.

One hypothesis is that rather than providing only an increase in hydrophobicity, the isoprenoid acts as part of a recognition unit for specific receptors that interact with either farnesylated or geranylgeranylated proteins. The recent observations that geranylgeranyl-modified forms of K-Ras4B or H-Ras proteins exhibit intracellular localizations which are different from those of their authentic farnesylated counterparts is consistent with this possibility.

Moreover, prenylation of nuclear lamins, which are involved in the mitotic control of membrane assembly, is necessary for the proper assembly of these proteins into the nuclear lamina. Indeed, prenylation is necessary to the maturation by cleavage of prelamin A in lamin A and to obtain functional lamin B.

Geranylgeranyl pyrophosphate synthetase (GGPS) is involved in the mevalonic acid biosynthetic pathway and is located in the cytosol. It catalyzes the consecutive condensation of isopentenyl diphosphate with allylic diphosphates to produce GGPP. This biosynthesis of GGPPS is regulated according to requirements for protein prenylation. GGPS has been found to be expressed in human fetal heart, as described in the PCT Application No WO 96/21736.

SUMMARY OF THE INVENTION

The present invention pertains to nucleic acid molecules comprising the genomic sequence of a novel human gene which encodes a hGGPPS protein. The *hGGPPS* genomic sequence comprises regulatory sequence located upstream (5'-end) and downstream (3'-end) of the transcribed portion of said gene, these regulatory sequences being also part of the invention.

The invention also deals with the complete sequence of two cDNAs encoding the hGGPPS protein, as well as with the corresponding translation product.

Oligonucleotide probes or primers hybridizing specifically with a *hGGPPS* genomic or cDNA sequences are also part of the present invention, as well as DNA amplification and detection methods using said primers and probes.

A further object of the invention consists of recombinant vectors comprising any of the nucleic acid sequences described above, and in particular of recombinant vectors comprising a

hGGPPS regulatory sequence or a sequence encoding a hGGPPS protein, as well as of cell hosts and transgenic non human animals comprising said nucleic acid sequences or recombinant vectors.

The invention also concerns a *hGGPPS*-related biallelic marker.

Finally, the invention is directed to methods for the screening of substances or molecules that modify or inhibit the expression of *hGGPPS*.

BRIEF DESCRIPTION OF THE DRAWING

Figure 1 : Map of the genomic, cDNA and coding (CDS) sequences of *hGGPS* : (1) upper line, genomic sequence; (2) cDNA sequence of SEQ ID No 2; (3) coding sequence (CDS).

Figure 2 : Map of the genomic, cDNA and coding (CDS) sequences of *hGGPS* : (1) upper line, genomic sequence; (2) cDNA sequence of SEQ ID No 3; (3) coding sequence (CDS).

Brief Description of the sequences provided in the Sequence Listing

SEQ ID No 1 contains a genomic sequence of *hGGPPS* comprising the 5' regulatory region (upstream untranscribed region), the exons and introns, and the 3' regulatory region (downstream untranscribed region).

SEQ ID No 2 contains a cDNA sequence of *hGGPPS* comprising the exons 1, 2, 3, and 4.

SEQ ID No 3 contains a cDNA sequence of *hGGPPS* comprising the exons 1bis, 2, 3, and 4.

SEQ ID No 4 contains the amino acid sequence encoded by the cDNA of SEQ ID No 2 or 3.

SEQ ID Nos 5 and 6 contain the fragments containing a polymorphic base of the biallelic marker 5-187-77.

SEQ ID No 7 contains the microsequencing primer of the biallelic marker 5-187-77.

SEQ ID Nos 8 and 9 contain the amplification primers of the biallelic marker 5-187-77.

SEQ ID No 10 contains a primer containing the additional PU 5' sequence described further in Example 3.

SEQ ID No 11 contains a primer containing the additional RP 5' sequence described further in Example 3.

DETAILED DESCRIPTION OF THE INVENTION

The *hGGPS* gene of the invention is located on chromosome 1, and more precisely on the 1q42-1q43 locus of this chromosome. This chromosome 1 locus has been shown to carry a predisposing gene for prostate cancer (Berthon et al., 1998).

The *hGGPS* gene of the invention is located in the vicinity of a retinoblastoma binding protein gene. Indeed, the coding sequence of this latter gene is on a strand which is opposite to the strand carrying the *hGGPS* Open Reading Frame.

The aim of the present invention is to provide polynucleotides derived from the *hGGPS* gene, particularly those useful to design suitable means for detecting the presence of this gene in a test sample or alternatively t discriminate between the *hGGPS* mRNA molecules that are present in

a test sample. Other polynucleotides of the invention are useful to design suitable means to express a desired polynucleotide of interest. The invention also relates to the hGGPS polypeptide having the amino acid sequence of SEQ ID No 4.

Definitions

Before describing the invention in greater detail, the following definitions are set forth to illustrate and define the meaning and scope of the terms used to describe the invention herein.

The term "*hGGPPS* gene", when used herein, encompasses mRNA and cDNA sequences encoding the hGGPPS protein. In the case of a genomic sequence, the *hGGPPS* gene also includes native regulatory regions which control the expression of the coding sequence of the *hGGPPS* gene.

The term "functionally active fragment" of the hGGPPS protein is intended to designate a polypeptide carrying at least one of the structural features of the hGGPPS protein involved in at least one of the biological functions and/or activity of the hGGPPS protein.

A "heterologous" or "exogenous" polynucleotide designates a purified or isolated nucleic acid that has been placed, by genetic engineering techniques, in the environment of unrelated nucleotide sequences, such as the final polynucleotide construct does not occur naturally. An illustrative, but not limitative, embodiment of such a polynucleotide construct may be represented by a polynucleotide comprising (1) a regulatory polynucleotide derived from the hGGPPS gene sequence and (2) a polynucleotide encoding a cytokine, for example GM-CSF. The polypeptide encoded by the heterologous polynucleotide will be termed an heterologous polypeptide for the purpose of the present invention.

By a "biologically active fragment or variant" of a regulatory polynucleotide according to the present invention is intended a polynucleotide comprising or alternatively consisting in a fragment of said polynucleotide which is functional as a regulatory region for expressing a recombinant polypeptide or a recombinant polynucleotide in a recombinant cell host.

For the purpose of the invention, a nucleic acid or polynucleotide is "functional" as a regulatory region for expressing a recombinant polypeptide or a recombinant polynucleotide if said regulatory polynucleotide contains nucleotide sequences which contain transcriptional and translational regulatory information, and such sequences are "operatively linked" to nucleotide sequences which encode the desired polypeptide or the desired polynucleotide. An operable linkage is a linkage in which the regulatory nucleic acid and the DNA sequence sought to be expressed are linked in such a way as to permit gene expression.

As used herein, the term "operably linked" refers to a linkage of polynucleotide elements in a functional relationship. For instance, a promoter or enhancer is operably linked to a coding sequence if it affects the transcription of the coding sequence. More precisely, two DNA molecules (such as a polynucleotide containing a promoter region and a polynucleotide encoding a desired polypeptide or polynucleotide) are said to be "operably linked" if the nature of the linkage between

the two polynucleotides does not (1) result in the introduction of a frame-shift mutation or (2) interfere with the ability of the polynucleotide containing the promoter to direct the transcription of the coding polynucleotide. The promoter polynucleotide would be operably linked to a polynucleotide encoding a desired polypeptide or a desired polynucleotide if the promoter is capable of effecting transcription of the polynucleotide of interest.

The terms "sample" or "material sample" are used herein to designate a solid or a liquid material suspected to contain a polynucleotide or a polypeptide of the invention. A solid material may be, for example, a tissue slice or biopsy within which is searched the presence of a polynucleotide encoding a hGGPPS protein, either a DNA or RNA molecule or within which is searched the presence of a native or a mutated hGGPPS protein, or alternatively the presence of a desired protein of interest the expression of which has been placed under the control of a *hGGPPS* regulatory polynucleotide. A liquid material may be, for example, any body fluid like serum, urine etc., or a liquid solution resulting from the extraction of nucleic acid or protein material of interest from a cell suspension or from cells in a tissue slice or biopsy. The term "biological sample" is also used and is more precisely defined within the Section dealing with DNA extraction.

As used herein, the term "purified" does not require absolute purity; rather, it is intended as a relative definition. Purification if starting material or natural material to at least one order of magnitude, preferably two or three orders, and more preferably four or five orders of magnitude is expressly contemplated. As an example, purification from 0.1% concentration to 10% concentration is two orders of magnitude.

The term "isolated" requires that the material be removed from its original environment (e.g. the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide or polypeptide present in a living animal is not isolated, but the same polynucleotide or DNA or polypeptide, separated from some or all of the coexisting materials in the natural system, is isolated. Such polynucleotide could be part of a vector and/or such polynucleotide or polypeptide could be part of a composition and still be isolated in that the vector or composition is not part of its natural environment.

The term "polypeptide" refers to a polymer of amino acids without regard to the length of the polymer; thus, peptides, oligopeptides, and proteins are included within the definition of polypeptide. This term also does not specify or exclude post-expression modifications of polypeptides, for example, polypeptides which include the covalent attachment of glycosyl groups, acetyl groups, phosphate groups, lipid groups and the like are expressly encompassed by the term polypeptide. Also included within the definition are polypeptides which contain one or more analogs of an amino acid (including, for example, non-naturally occurring amino acids, amino acids which only occur naturally in an unrelated biological system, modified amino acids from mammalian systems etc.), polypeptides with substituted linkages, as well as other modifications known in the art, both naturally occurring and non-naturally occurring.

The term "recombinant polypeptide" is used herein to refer to polypeptides that have been artificially designed and which comprise at least two polypeptide sequences that are not found as contiguous polypeptide sequences in their initial natural environment, or to refer to polypeptides which have been expressed from a recombinant polynucleotide.

The term "purified" is used herein to describe a polypeptide of the invention which has been separated from other compounds including, but not limited to nucleic acids, lipids, carbohydrates and other proteins. A polypeptide is substantially pure when at least about 50%, preferably 60 to 75% of a sample exhibits a single polypeptide sequence. A substantially pure polypeptide typically comprises about 50%, preferably 60 to 90% weight/weight of a protein sample, more usually about 95%, and preferably is over about 99% pure. Polypeptide purity or homogeneity is indicated by a number of means well known in the art, such as polyacrylamide gel electrophoresis of a sample, followed by visualizing a single polypeptide band upon staining the gel. For certain purposes higher resolution can be provided by using HPLC or other means well known in the art.

As used herein, the term "non-human animal" refers to any non-human vertebrate, birds and more usually mammals, preferably primates, farm animals such as swine, goats, sheep, donkeys, and horses, rabbits or rodents, more preferably rats or mice. As used herein, the term "animal" is used to refer to any vertebrate, preferable a mammal. Both the terms "animal" and "mammal" expressly embrace human subjects unless preceded with the term "non-human".

As used herein, the term "antibody" refers to a polypeptide or group of polypeptides which are comprised of at least one binding domain, where an antibody binding domain is formed from the folding of variable domains of an antibody molecule to form three-dimensional binding spaces with an internal surface shape and charge distribution complementary to the features of an antigenic determinant of an antigen, which allows an immunological reaction with the antigen. Antibodies include recombinant proteins comprising the binding domains, as wells as fragments, including Fab, Fab', $F(ab)_2$, and $F(ab')_2$ fragments.

As used herein, an "antigenic determinant" is the portion of an antigen molecule, in this case a hGGPPS polypeptide, that determines the specificity of the antigen-antibody reaction. An "epitope" refers to an antigenic determinant of a polypeptide. An epitope can comprise as few as 3 amino acids in a spatial conformation which is unique to the epitope. Generally an epitope consists of at least 6 such amino acids, and more usually at least 8-10 such amino acids. Methods for determining the amino acids which make up an epitope include x-ray crystallography, 2-dimensional nuclear magnetic resonance, and epitope mapping e.g. the Pepscan method described by Geysen et al. 1984; PCT Publication No. WO 84/03564; and PCT Publication No. WO 84/03506.

Throughout the present specification, the expression "nucleotide sequence" may be employed to designate indifferently a polynucleotide or an oligonucleotide or a nucleic acid. More precisely, the expression "nucleotide sequence" encompasses the nucleic material itself and is thus

not restricted to the sequence information (i.e. the succession of letters chosen among the four base letters) that biochemically characterizes a specific DNA or RNA molecule.

As used interchangeably herein, the term "oligonucleotides", and "polynucleotides" include RNA, DNA, or RNA/DNA hybrid sequences of more than one nucleotide in either single chain or duplex form. The term "nucleotide" as used herein as an adjective to describe molecules comprising RNA, DNA, or RNA/DNA hybrid sequences of any length in single-stranded or duplex form. The term "nucleotide" is also used herein as a noun to refer to individual nucleotides or varieties of nucleotides, meaning a molecule, or individual unit in a larger nucleic acid molecule, comprising a purine or pyrimidine, a ribose or deoxyribose sugar moiety, and a phosphate group, or phosphodiester linkage in the case of nucleotides within an oligonucleotide or polynucleotide. Although the term "nucleotide" is also used herein to encompass "modified nucleotides" which comprise at least one modifications (a) an alternative linking group, (b) an analogous form of purine, (c) an analogous form of pyrimidine, or (d) an analogous sugar, for examples of analogous linking groups, purine, pyrimidines, and sugars see for example PCT publication No WO 95/04064. However, the polynucleotides of the invention are preferably comprised of greater than 50% conventional deoxyribose nucleotides, and most preferably greater than 90% conventional deoxyribose nucleotides. The polynucleotide sequences of the invention may be prepared by any known method, including synthetic, recombinant, *ex vivo* generation, or a combination thereof, as well as utilizing any purification methods known in the art.

The term "heterozygosity rate" is used herein to refer to the incidence of individuals in a population which are heterozygous at a particular allele. In a biallelic system, the heterozygosity rate is on average equal to $2P_a(1-P_a)$, where P_a is the frequency of the least common allele. In order to be useful in genetic studies, a genetic marker should have an adequate level of heterozygosity to allow a reasonable probability that a randomly selected person will be heterozygous.

The term "genotype" as used herein refers the identity of the alleles present in an individual or a sample. In the context of the present invention a genotype preferably refers to the description of the biallelic marker alleles present in an individual or a sample. The term "genotyping" a sample or an individual for a biallelic marker consists of determining the specific allele or the specific nucleotide carried by an individual at a biallelic marker.

The term "polymorphism" as used herein refers to the occurrence of two or more alternative genomic sequences or alleles between or among different genomes or individuals. "Polymorphic" refers to the condition in which two or more variants of a specific genomic sequence can be found in a population. A "polymorphic site" is the locus at which the variation occurs. A single nucleotide polymorphism is a single base pair change. Typically a single nucleotide polymorphism is the replacement of one nucleotide by another nucleotide at the polymorphic site. Deletion of a single nucleotide or insertion of a single nucleotide, also give rise to single nucleotide polymorphisms. In the context of the present invention "single nucleotide polymorphism" preferably refers to a single

nucleotide substitution. Typically, between different genomes or between different individuals, the polymorphic site may be occupied by two different nucleotides.

The term "biallelic polymorphism" and "biallelic marker" are used interchangeably herein to refer to a single nucleotide polymorphism having two alleles at a fairly high frequency in the population. A "biallelic marker allele" refers to the nucleotide variants present at a biallelic marker site. Typically, the frequency of the less common allele of the biallelic markers of the present invention has been validated to be greater than 1%, preferably the frequency is greater than 10%, more preferably the frequency is at least 20% (i.e. heterozygosity rate of at least 0.32), even more preferably the frequency is at least 30% (i.e. heterozygosity rate of at least 0.42). A biallelic marker wherein the frequency of the less common allele is 30% or more is termed a "high quality biallelic marker".

The location of nucleotides in a polynucleotide with respect to the center of the polynucleotide are described herein in the following manner. When a polynucleotide has an odd number of nucleotides, the nucleotide at an equal distance from the 3' and 5' ends of the polynucleotide is considered to be "at the center" of the polynucleotide, and any nucleotide immediately adjacent to the nucleotide at the center, or the nucleotide at the center itself is considered to be "within 1 nucleotide of the center." With an odd number of nucleotides in a polynucleotide any of the five nucleotides positions in the middle of the polynucleotide would be considered to be within 2 nucleotides of the center, and so on. When a polynucleotide has an even number of nucleotides, there would be a bond and not a nucleotide at the center of the polynucleotide. Thus, either of the two central nucleotides would be considered to be "within 1 nucleotide of the center" and any of the four nucleotides in the middle of the polynucleotide would be considered to be "within 2 nucleotides of the center", and so on.

As used herein the terminology "defining a biallelic marker" means that a sequence includes a polymorphic base from a biallelic marker. The sequences defining a biallelic marker may be of any length consistent with their intended use, provided that they contain a polymorphic base from a biallelic marker. The sequence has between 1 and 500 nucleotides in length, preferably between 5, 10 , 15, 20, 25, or 40 and 200 nucleotides and more preferably between 30 and 50 nucleotides in length. Each biallelic marker therefore corresponds to two forms of a polynucleotide sequence included in a gene, which, when compared with one another, present a nucleotide modification at one position. Preferably, the sequences defining a biallelic marker include a polymorphic base of the biallelic marker 5-187-77. In some embodiments the sequences defining a biallelic marker comprise one of the sequences selected from the group consisting of SEQ ID Nos 5 and 6. Likewise, the term "marker" or "biallelic marker" requires that the sequence is of sufficient length to practically (although not necessarily unambiguously) identify the polymorphic allele, which usually implies a length of at least 4, 5, 6, 10, 15, 20, 25, or 40 nucleotides.

The terms "base paired" and "Watson & Crick base paired" are used interchangeably herein to refer to nucleotides which can be hydrogen bonded to one another be virtue of their sequence identities in a manner like that found in double-helical DNA with thymine or uracil residues linked to adenine residues by two hydrogen bonds and cytosine and guanine residues linked by three hydrogen bonds (See Stryer, L., *Biochemistry*, 4th edition, 1995).

The terms "complementary" or "complement thereof" are used herein to refer to the sequences of polynucleotides which is capable of forming Watson & Crick base pairing with another specified polynucleotide throughout the entirety of the complementary region. For the purpose of the present invention, a first polynucleotide is deemed to be complementary to a second polynucleotide when each base in the first polynucleotide is paired with its complementary base. Complementary bases are, generally, A and T (or A and U), or C and G. "Complement" is used herein as a synonym from "complementary polynucleotide", "complementary nucleic acid" and "complementary nucleotide sequence". These terms are applied to pairs of polynucleotides based solely upon their sequences and not any particular set of conditions under which the two polynucleotides would actually bind.

Variants and fragments

1. Polynucleotides

The invention also relates to variants and fragments of the polynucleotides described herein, particularly of a *hGGPPS* gene containing one or more biallelic markers according to the invention.

Variants of polynucleotides, as the term is used herein, are polynucleotides that differ from a reference polynucleotide. A variant of a polynucleotide may be a naturally occurring variant such as a naturally occurring allelic variant, or it may be a variant that is not known to occur naturally. Such non-naturally occurring variants of the polynucleotide may be made by mutagenesis techniques, including those applied to polynucleotides, cells or organisms. Generally, differences are limited so that the nucleotide sequences of the reference and the variant are closely similar overall and, in many regions, identical.

Variants of polynucleotides according to the invention include, without being limited to, nucleotide sequences that are at least 95% identical to any of SEQ ID Nos 1-3 or the sequences complementary thereto or to any polynucleotide fragment of at least 8 consecutive nucleotides of any of SEQ ID Nos 1-3 or the sequences complementary thereto, and preferably at least 98% identical, more particularly at least 99.5% identical, and most preferably at least 99.9% identical to any of SEQ ID Nos 1-3 or the sequences complementary thereto or to any polynucleotide fragment of at least 8 consecutive nucleotides of any of SEQ ID Nos 1-3 or the sequences complementary thereto.

Changes in the nucleotide of a variant may be silent, which means that they do not alter the amino acids encoded by the polynucleotide.

However, nucleotide changes may also result in amino acid substitutions, additions, deletions, fusions and truncations in the polypeptide encoded by the reference sequence. The substitutions, deletions or additions may involve one or more nucleotides. The variants may be altered in coding or non-coding regions or both. Alterations in the coding regions may produce conservative or non-conservative amino acid substitutions, deletions or additions.

In the context of the present invention, particularly preferred embodiments are those in which the polynucleotides encode polypeptides which retain substantially the same biological function or activity as the mature hGGPPS protein.

A polynucleotide fragment is a polynucleotide having a sequence that entirely is the same as part but not all of a given nucleotide sequence, preferably the nucleotide sequence of a *hGGPPS* gene, and variants thereof. The fragment can be a portion of an exon or of an intron of a *hGGPPS* gene. It can also be a portion of the regulatory sequences of the *hGGPPS* gene. Preferably, such fragments comprise the polymorphic base of the biallelic marker 5-187-77 of SEQ ID Nos 5-6.

Such fragments may be "free-standing", i.e. not part of or fused to other polynucleotides, or they may be comprised within a single larger polynucleotide of which they form a part or region. However, several fragments may be comprised within a single larger polynucleotide.

As representative examples of polynucleotide fragments of the invention, there may be mentioned those which have from about 4, 6, 8, 15, 20, 25, 40, 10 to 20, 10 to 30, 30 to 55, 50 to 100, 75 to 100 or 100 to 200 nucleotides in length. Preferred are those fragments having about 49 nucleotides in length, such as those of SEQ ID Nos 5-6 or the sequences complementary thereto and containing at least one of the biallelic markers of a *hGGPPS* gene which are described herein.

2. Polypeptides.

The invention also relates to variants, fragments, analogs and derivatives of the polypeptides described herein, including mutated hGGPPS proteins.

The variant may be 1) one in which one or more of the amino acid residues are substituted with a conserved or non-conserved amino acid residue (preferably a conserved amino acid residue) and such substituted amino acid residue may or may not be one encoded by the genetic code, or 2) one in which one or more of the amino acid residues includes a substituent group, or 3) one in which the mutated hGGPPS is fused with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol), or 4) one in which the additional amino acids are fused to the mutated hGGPPS, such as a leader or secretory sequence or a sequence which is employed for purification of the mutated hGGPPS or a preprotein sequence. Such variants are deemed to be within the scope of those skilled in the art.

More particularly, a variant hGGPPS polypeptide comprises amino acid changes ranging from 1, 2, 3, 4, 5, 10 to 20 substitutions, additions or deletions of one aminoacid, preferably from 1 to 10, more preferably from 1 to 5 and most preferably from 1 to 3 substitutions, additions or deletions of one amino acid. The preferred amino acid changes are those which have little or no

influence on the biological activity or the capacity of the variant hGGPPS polypeptide to be recognized by antibodies raised against a native hGGPPS protein.

By homologous peptide according to the present invention is meant a polypeptide containing one or several aminoacid additions, deletions and/or substitutions in the amino acid sequence of a hGGPPS polypeptide. In the case of an aminoacid substitution, one or several -consecutive or non-consecutive- aminoacids are replaced by « equivalent » aminoacids.

The expression "equivalent" amino acid is used herein to designate any amino acid that may be substituted for one of the amino acids having similar properties, such that one skilled in the art of peptide chemistry would expect the secondary structure and hydropathic nature of the polypeptide to be substantially unchanged. Generally, the following groups of amino acids represent equivalent changes: (1) Ala, Pro, Gly, Glu, Asp, Gln, Asn, Ser, Thr; (2) Cys, Ser, Tyr, Thr; (3) Val, Ile, Leu, Met, Ala, Phe; (4) Lys, Arg, His; (5) Phe, Tyr, Trp, His.

By an equivalent aminoacid according to the present invention is also meant the replacement of a residue in the L-form by a residue in the D form or the replacement of a Glutamic acid (E) residue by a Pyro-glutamic acid compound. The synthesis of peptides containing at least one residue in the D-form is, for example, described by Koch (1977).

A specific, but not restrictive, embodiment of a modified peptide molecule of interest according to the present invention, which consists in a peptide molecule which is resistant to proteolysis, is a peptide in which the -CONH- peptide bond is modified and replaced by a (CH_2NH) reduced bond, a (NHCO) retro inverso bond, a (CH_2-O) methylene-oxy bond, a (CH_2-S) thiomethylene bond, a (CH_2CH_2) carba bond, a (CO-CH_2) cetomethylene bond, a (CHOH-CH_2) hydroxyethylene bond), a (N-N) bound, a E-alcene bond or also a -CH=CH- bond.

The polypeptide accoding to the invention could have post-translational modifications. For example, it can present the following modifications: acylation, disulfide bond formation, prenylation, carboxymethylation and phosphorylation.

A polypeptide fragment is a polypeptide having a sequence that entirely is the same as part but not all of a given polypeptide sequence, preferably a polypeptide encoded by a *hGGPPS* gene and variants thereof. Preferred fragments include those regions possessing antigenic properties and which can be used to raise antibodies against the hGGPPS protein.

Such fragments may be "free-standing", i.e. not part of or fused to other polypeptides, or they may be comprised within a single larger polypeptide of which they form a part or region. However, several fragments may be comprised within a single larger polypeptide.

As representative examples of polypeptide fragments of the invention, there may be mentioned those which comprise at least about 5, 6, 7, 8, 9 or 10 to 15, 10 to 20, 15 to 40, or 30 to 55 amino acids of the hGGPPS. In some embodiments, the fragments contain at least one amino acid mutation in the hGGPPS protein.

Identity Between Nucleic Acids Or Polypeptides

The terms "percentage of sequence identity" and "percentage homology" are used interchangeably herein to refer to comparisons among polynucleotides and polypeptides, and are determined by comparing two optimally aligned sequences over a comparison window, wherein the portion of the polynucleotide or polypeptide sequence in the comparison window may comprise additions or deletions (i.e., gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical nucleic acid base or amino acid residue occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence identity. Homology is evaluated using any of the variety of sequence comparison algorithms and programs known in the art. Such algorithms and programs include, but are by no means limited to, TBLASTN, BLASTP, FASTA, TFASTA, and CLUSTALW (Pearson and Lipman, 1988; Altschul et al., 1990; Thompson et al., 1994; Higgins et al., 1996; Altschul et al., 1993). In a particularly preferred embodiment, protein and nucleic acid sequence homologies are evaluated using the Basic Local Alignment Search Tool ("BLAST") which is well known in the art (see, e.g., Karlin and Altschul, 1990; Altschul et al., 1990, 1993, 1997). In particular, five specific BLAST programs are used to perform the following task:

(1) BLASTP and BLAST3 compare an amino acid query sequence against a protein sequence database;

(2) BLASTN compares a nucleotide query sequence against a nucleotide sequence database;

(3) BLASTX compares the six-frame conceptual translation products of a query nucleotide sequence (both strands) against a protein sequence database;

(4) TBLASTN compares a query protein sequence against a nucleotide sequence database translated in all six reading frames (both strands); and

(5) TBLASTX compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

The BLAST programs identify homologous sequences by identifying similar segments, which are referred to herein as "high-scoring segment pairs," between a query amino or nucleic acid sequence and a test sequence which is preferably obtained from a protein or nucleic acid sequence database. High-scoring segment pairs are preferably identified (i.e., aligned) by means of a scoring matrix, many of which are known in the art. Preferably, the scoring matrix used is the BLOSUM62 matrix (Gonnet et al., 1992; Henikoff and Henikoff, 1993). Less preferably, the PAM or PAM250 matrices may also be used (see, e.g., Schwartz and Dayhoff, eds., 1978). The BLAST programs evaluate the statistical significance of all high-scoring segment pairs identified, and preferably selects those segments which satisfy a user-specified threshold of significance, such as a user-

specified percent homology. Preferably, the statistical significance of a high-scoring segment pair is evaluated using the statistical significance formula of Karlin (see, e.g., Karlin and Altschul, 1990).

Stringent Hybridization Conditions

By way of example and not limitation, procedures using conditions of high stringency are as follows: Prehybridization of filters containing DNA is carried out for 8 h to overnight at 65°C in buffer composed of 6X SSC, 50 mM Tris-HCl (pH 7.5), 1 mM EDTA, 0.02% PVP, 0.02% Ficoll, 0.02% BSA, and 500 µg/ml denatured salmon sperm DNA. Filters are hybridized for 48 h at 65°C, the preferred hybridization temperature, in prehybridization mixture containing 100 µg/ml denatured salmon sperm DNA and 5-20 X 10^6 cpm of ^{32}P-labeled probe. Alternatively, the hybridization step can be performed at 65°C in the presence of SSC buffer, 1 x SSC corresponding to 0.15M NaCl and 0.05 M Na citrate. Subsequently, filter washes can be done at 37°C for 1 h in a solution containing 2 x SSC, 0.01% PVP, 0.01% Ficoll, and 0.01% BSA, followed by a wash in 0.1 X SSC at 50°C for 45 min. Alternatively, filter washes can be performed in a solution containing 2 x SSC and 0.1% SDS, or 0.5 x SSC and 0.1% SDS, or 0.1 x SSC and 0.1% SDS at 68°C for 15 minute intervals. Following the wash steps, the hybridized probes are detectable by autoradiography. Other conditions of high stringency which may be used are well known in the art and as cited in Sambrook et al., 1989; and Ausubel et al., 1989, are incorporated herein in their entirety. These hybridization conditions are suitable for a nucleic acid molecule of about 20 nucleotides in length. There is no need to say that the hybridization conditions described above are to be adapted according to the length of the desired nucleic acid, following techniques well known to the one skilled in the art. The suitable hybridization conditions may for example be adapted according to the teachings disclosed in the book of Hames and Higgins (1985) or in Sambrook et al.(1989).

hGGPS gene polynucleotide, cDNAs and associated regulatory regions.

Genomic sequences

The invention concerns a purified or isolated nucleic acid encoding the hGGPS polypeptide, wherein said nucleic acid comprises the nucleotide sequence of SEQ ID No 1.

The present invention concerns a purified or isolated nucleic acid comprising a nucleotide sequence of SEQ ID No 1, or a nucleotide sequence complementary thereto or a fragment or a variant thereof.

Particularly preferred nucleic acids of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of SEQ ID No 1 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of the following nucleotide positions of SEQ ID No 1: 1-485, 547-632, 827-7291, 7385-13759, 13831-14062, 14671-15054, and 15252-17131.

The invention also encompasses a purified or isolated nucleic acid having at least 95% nucleotide identity with the nucleotide sequence of SEQ ID No 1 or a complementary sequence thereto.

A further object of the invention consists in a purified or isolated nucleic acid of at least 12 nucleotides in length, wherein said nucleic acid hybridizes under stringent hybridization conditions with a polynucleotide sequence of SEQ ID No 1 or a complementary sequence thereto.

The *hGGPS* genomic nucleic acid sequence comprises five exons. These five exons are described in Table A.

Table A

Exon	Beginning position in SEQ ID No 1	End position In SEQ ID No 1	Intron	Beginning position in SEQ ID No 1	End position In SEQ ID No 1
1	486	546	1	547	7291
1bis	633	826	1bis	827	7291
2	7292	7384	2	7385	13759
3	13760	13830	3	13831	14062
4	14063	15251			

The hGGPS introns defined hereinafter for the purpose of the present invention are not exactly what is generally understood as "introns" by the one skilled in the art and will consequently be defined below.

Generally, an intron is defined as a nucleotide sequence that is present both in the genomic DNA and in the unspliced mRNA molecule, and which is absent from the mRNA molecule which has undergone the splicing events. In the case of the *hGGPS* gene, the inventors have found that at least two different spliced mRNA molecules are produced when this gene is transcribed, as it will be described in detail in a further section of the specification. The first spliced mRNA molecule comprises Exons 1, 2, 3 and 4, as shown in Figure 1. Thus, the genomic nucleotide sequence comprised between Exon 1 and Exon 2 is an intronic sequence as regards to this first mRNA molecule, despite the fact that this intronic sequence contains Exon 1*bis*. In contrast, Exon 1*bis* is of course an exonic nucleotide sequence as regards to the second *hGGPS* mRNA molecule shown in Figure 2.

For the purpose of the present invention and in order to make a clear and unique designation of the different nucleic acids of the invention, it has been postulated that the polynucleotides contained both in the nucleotide sequence of SEQ ID No 1 and in any of the nucleotide sequences of SEQ ID Nos 2 or 3 are considered as exonic sequences. Conversely, the polynucleotides contained in the nucleotide sequence of SEQ ID No 1 and located between Exon 1 and Exon 4, but which are absent both from the nucleotide sequence of SEQ ID No 2 and from the nucleotide sequence of SEQ ID No 3 are considered as intronic sequences.

Thus, the invention embodies purified, isolated, or recombinant polynucleotides comprising a nucleotide sequence selected from the group consisting of the exons of the *hGGPPS* gene, or a sequence complementary thereto. The invention also deals with purified, isolated, or recombinant

nucleic acids comprising a combination of at least two exons of the *hGGPPS* gene, wherein the polynucleotides are arranged within the nucleic acid, from the 5'-end to the 3'-end of said nucleic acid, in the same order as in SEQ ID No 1.

The nucleic acids defining the *hGGPS* introns described above, as well as their fragments and variants, may be used as oligonucleotide primers or probes in order to detect the presence of a copy of the *hGGPS* in a test sample, or alternatively in order to amplify a target nucleotide sequence within the *hGGPS* intronic sequences.

***hGGPS* cDNAs**

The inventors have discovered that the expression of the *hGGPS* gene leads to the production of at least two mRNA molecules, respectively a first and a second *hGGPS* transcription product.

The first transcription product comprises Exons 1, 2, 3 and 4. This cDNA of SEQ ID No 2 includes a 5'-UTR region, spanning the whole Exon 1 and part of Exon 2. This 5'-UTR region starts from the nucleotide at position 1 and ends at the nucleotide in position 84 of SEQ ID No 2. The cDNA of SEQ ID No 2 includes a 3'-UTR region starting from the nucleotide at position 988 and ending at the nucleotide at position 1414 of SEQ ID No 2. The 3'UTR carries a potential polyadenylation signal located between the nucleotide in position 1289 and the nucleotide in position 1294 of the nucleic acid of SEQ ID No 2. The ORF encoding hGGPS is comprised between the nucleotide in position 85 and the nucleotide in position 987 of SEQ ID No 2.

The second transcription product comprises Exons *1bis*, *2*, 3 and 4. This cDNA of SEQ ID No 3 includes a 5'-UTR region starting from the nucleotide at position 1 and ending at the nucleotide in position 217 of SEQ ID No 3. The cDNA of SEQ ID No 3 includes a 3'-UTR region starting from the nucleotide at position 1121 and ending at the nucleotide at position 1547 of SEQ ID No 3. The 3'UTR carries a potential polyadenylation signal located between the nucleotide in position 1422 and the nucleotide in position 1427 of the nucleic acid of SEQ ID No 3. The ORF encoding hGGPS is comprised between the nucleotide in position 218 and the nucleotide in position 1120 of the nucleotide sequence of SEQ ID No 3.

Another object of the invention consists of a purified or isolated nucleic acid selected from the group consisting of the nucleotide sequences of SEQ ID Nos 2 and 3 or a complementary sequence thereto or a fragment thereof.

Particularly preferred nucleic acids of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of SEQ ID No 2 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of the nucleotide positions 834-1217 of SEQ ID No 2. Additional preferred nucleic acids of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of SEQ ID No 2 or the

complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of the nucleotide positions 967-1351 of SEQ ID No 3.

The invention also pertains to a purified or isolated nucleic acid having at least 95% of nucleotide identity with any of the nucleotide sequences of SEQ ID Nos 2 and 3 or a complementary sequence thereto.

A further object of the invention consists in a purified or isolated nucleic acid of at least 12 nucleotides in length, wherein said nucleic acid hybridizes under stringent hybridization conditions with a polynucleotide selected from the group consisting of the nucleotide sequences of SEQ ID Nos 2 and 3, or a sequence complementary thereto.

Another object of the invention consists in a purified or isolated nucleic acid comprising a nucleic acid fragment of a nucleotide sequence selected from the group consisting of SEQ ID Nos 2 and 3, wherein this nucleic acid fragment encodes a polypeptide having an amino acid sequence beginning at the amino acid in position 200 and ending at the amino acid in position 300 of the hGGPS polypeptide of SEQ ID No 4, or a nucleic acid encoding a peptide fragment thereof.

Regulatory sequences

As already mentioned hereinbefore, the polynucleotide of SEQ ID No 1 contains regulatory sequences both in the non-coding 5'-flanking region and in the non-coding 3'-flanking region that border the *hGGPS* coding region.

The longest 5'-regulatory sequence of the *hGGPS* gene is localized between the nucleotide in position 1 and the nucleotide in position 632 of SEQ ID No1. However, a shorter 5'-regulatory sequence of the *hGGPS* gene is localized between the nucleotide in position 1 and the nucleotide in position 485 of SEQ ID No1.

The *hGGPS* 3'-regulatory region, as shown in Figure 1, comprises a nucleotide sequence starting from the nucleotide in position 15252 of SEQ ID No 1 and ending at the nucleotide in position 17131 of SEQ ID No1.

Polynucleotides derived from the *hGGPS* regulatory regions described above are useful in order to detect the presence of at least a copy of the nucleotide sequence of SEQ ID No 1 in a test sample.

The promoter activity of the regulatory regions contained in the *hGGPS* nucleotide sequence of SEQ ID No 1 can be assessed as described below.

Genomic sequences located upstream of the *hGGPS* gene are cloned into a suitable promoter reporter vector, such as the pSEAP-Basic, pSEAP-Enhancer, pβgal-Basic, pβgal-Enhancer, or pEGFP-1 Promoter Reporter vectors available from Clontech. Briefly, each of these promoter reporter vectors include multiple cloning sites positioned upstream of a reporter gene encoding a readily assayable protein such as secreted alkaline phosphatase, beta galactosidase, or green fluorescent protein. The sequences upstream the *hGGPS* coding region are inserted into the cloning sites upstream of the reporter gene in both orientations and introduced into an appropriate host cell.

The level of reporter protein is assayed and compared to the level obtained from a vector which lacks an insert in the cloning site. The presence of an elevated expression level in the vector containing the insert with respect to the control vector indicates the presence of a promoter in the insert. If necessary, the upstream sequences can be cloned into vectors which contain an enhancer for increasing transcription levels from weak promoter sequences. A significant level of expression above that observed with the vector lacking an insert indicates that a promoter sequence is present in the inserted upstream sequence.

Promoter sequences within the upstream genomic DNA may be further defined by constructing nested deletions in the upstream DNA using conventional techniques such as Exonuclease III digestion. The resulting deletion fragments can be inserted into the promoter reporter vector to determine whether the deletion has reduced or obliterated promoter activity. In this way, the boundaries of the promoters may be defined. If desired, potential individual regulatory sites within the promoter may be identified using site directed mutagenesis or linker scanning to obliterate potential transcription factor binding sites within the promoter individually or in combination. The effects of these mutations on transcription levels may be determined by inserting the mutations into cloning sites in promoter reporter vectors.

Polynucleotides carrying the regulatory elements located both at the 5' end and at the 3' end of the *hGGPS* coding region may be advantageously used to control the transcriptional and translational activity of an heterologous polynucleotide of interest.

Thus, the present invention also concerns a purified or isolated nucleic acid comprising a polynucleotide which is selected from the group consisting of the 5' and 3' regulatory regions, or a sequence complementary thereto or a biologically active fragment or variant thereof. "5' regulatory region" refers to the nucleotide sequence located between positions 1 and 632 of SEQ ID No 1. "3' regulatory region" refers to the nucleotide sequence located between positions 15252 and 17131 of SEQ ID No 1.

The present invention is also directed to a polynucleotide comprising a functional portion of a regulatory region contained in the contemplated *hGGPS* gene and to its use in a recombinant expression vector carrying a polynucleotide encoding a polypeptide or a nucleic acid of interest.

Preferred fragments of the 5' regulatory region have a length of about 400 nucleotides, more particularly about 300 nucleotides, more preferably 200 nucleotides and most preferably about 100 nucleotides.

Preferred fragments of the 3' regulatory region have a length of about 600 nucleotides, more particularly about 300 nucleotides, more preferably 200 nucleotides and most preferably about 100 nucleotides.

In order, to identify the relevant biologically active polynucleotide derivatives of the 5' and 3' regulatory regions, the one skill in the art will refer to the book of Sambrook et al. (1989) which describes the use of a recombinant vector carrying a marker gene (i.e. beta galactosidase.

chloramphenicol acetyl transferase, etc.) the expression of which will be detected when placed under the control of a biologically active derivative polynucleotide of the 5' and 3' regulatory regions.

The regulatory polynucleotides of the invention may be prepared from a polynucleotide of the nucleotide sequence SEQ ID No 1 by cleavage using suitable restriction enzymes, as described for example in the book of Sambrook et al. (1989). The regulatory polynucleotides may also be prepared by digestion of a polynucleotide of the nucleotide sequence SEQ ID No 1 by an exonuclease enzyme, such as for example Bal31 (Wabiko et al., 1986). These regulatory polynucleotides can also be prepared by nucleic acid chemical synthesis, as described elsewhere in the specification.

The regulatory polynucleotides according to the invention may be advantageously part of a recombinant expression vector that may be used to express a coding sequence in a desired host cell or host organism. The recombinant expression vectors according to the invention are described elsewhere in the specification.

A preferred 5'-regulatory polynucleotide of the invention includes the 5'-untranslated region (5'-UTR) located between the nucleotide at position 1 and the nucleotide at position 84 of SEQ ID No 2, or a biologically active fragment or variant thereof.

Another preferred 5'-regulatory polynucleotide of the invention includes the 5'-untranslated region (5'-UTR) located between the nucleotide at position 1 and the nucleotide at position 217 of SEQ ID No 3, or a biologically active fragment or variant thereof.

A preferred 3'-regulatory polynucleotide of the invention includes the 3'-untranslated region (3'-UTR) consisting in the nucleotide sequence starting from the nucleotide in position 988 and ending a the nucleotide in position 1414 of the nucleic acid of SEQ ID No 2.

A further object of the invention consists of a purified or isolated nucleic acid comprising :

a) a nucleic acid comprising the 5' regulatory region or a biologically active fragment or variant thereof;

b) a polynucleotide encoding a desired polypeptide or nucleic acid operably linked to the 5' regulatory region or its biologically active fragment or variant thereof;

c) optionally, a nucleic acid comprising the 3' regulatory region or a biologically active fragment or variant thereof.

The desired polypeptide encoded by the above described nucleic acid may be of various nature or origin, encompassing proteins of prokaryotic or eukaryotic origin. Among the polypeptides expressed under the control of a *hGGPS* regulatory region, there may be cited bacterial, fungal or viral antigens. Also encompassed are eukaryotic proteins such as intracellular proteins, like "house keeping" proteins, membrane-bound proteins, like receptors, and secreted proteins like the numerous endogenous mediators such as cytokines.

The desired nucleic acids encoded by the above described polynucleotide, usually a RNA molecule, may be complementary to a desired coding polynucleotide, for example to the *hGGPS* coding sequence, and thus useful as an antisense polynucleotide.

Such a polynucleotide may be included in a recombinant expression vector in order to express the desired polypeptide or the desired nucleic acid in host cell or in a host organism. Suitable recombinant vectors that contain a polynucleotide such as described hereinbefore are disclosed elsewhere in the specification.

Coding regions

The *hGGPS* open reading frame is contained in the corresponding mRNAs of SEQ ID Nos 2 and 3.

More precisely, the effective *hGGPS* coding sequence (CDS) is comprised between the nucleotide at position 85 (first nucleotide of the ATG codon) and the nucleotide at position 987 (end nucleotide of the TAA codon) of SEQ ID No 2. A purified or isolated polynucleotide comprising the *hGGPS* coding region defined above is another object of the invention.

The above disclosed polynucleotide that contains the coding sequence of the *hGGPS* gene of the invention may be expressed in a desired host cell or a desired host organism, when this polynucleotide is placed under the control of suitable expression signals. The expression signals may be either the expression signals contained in the regulatory regions in the *hGGPS* gene of the invention or in contrast be exogenous regulatory nucleic sequences. Such a polynucleotide, when placed under the suitable expression signals, may also be inserted in a vector for its expression.

Biallelic Markers

The inventors have discovered nucleotide polymorphisms located within the genomic DNA containing the *hGGPS* gene, and among them SNP that are also termed biallelic markers. The biallelic markers of the invention can be used for example for the generation of genetic map, the linkage analysis, the association studies.

A) Identification Of Biallelic Markers

There are two preferred methods through which the biallelic markers of the present invention can be generated. In a first method, DNA samples from unrelated individuals are pooled together, following which the genomic DNA of interest is amplified and sequenced. The nucleotide sequences thus obtained are then analyzed to identify significant polymorphisms.

One of the major advantages of this method resides in the fact that the pooling of the DNA samples substantially reduces the number of DNA amplification reactions and sequencing reactions which must be carried out. Moreover, this method is sufficiently sensitive so that a biallelic marker obtained therewith usually shows a sufficient degree of informativeness for conducting association studies.

In a second method for generating biallelic markers, the DNA samples are not pooled and are therefore amplified and sequenced individually. The resulting nucleotide sequences obtained are then also analyzed to identify significant polymorphisms.

The following is a description of the various parameters of a preferred method used by the inventors to generate the markers of the present invention.

1. DNA extraction

The genomic DNA samples from which the biallelic markers of the present invention are generated are preferably obtained from unrelated individuals corresponding to a heterogeneous population of known ethnic background.

The number of individuals from whom DNA samples are obtained can vary substantially, preferably from about 10 to about 1000, preferably from about 50 to about 200 individuals. It is usually preferred to collect DNA samples from at least about 100 individuals in order to have sufficient polymorphic diversity in a given population to identify as many markers as possible and to generate statistically significant results.

As for the source of the genomic DNA to be subjected to analysis, any test sample can be foreseen without any particular limitation. These test samples include biological samples which can be tested by the methods of the present invention described herein and include human and animal body fluids such as whole blood, serum, plasma, cerebrospinal fluid, urine, lymph fluids, and various external secretions of the respiratory, intestinal and genitourinary tracts, tears, saliva, milk, white blood cells, myelomas and the like; biological fluids such as cell culture supernatants; fixed tissue specimens including tumor and non-tumor tissue and lymph node tissues; bone marrow aspirates and fixed cell specimens. The preferred source of genomic DNA used in the context of the present invention is from peripheral venous blood of each donor.

The techniques of DNA extraction are well-known to the skilled technician. Such techniques are described notably by Lin et al. (1998) and by Mackey et al. (1998). Details of a preferred embodiment are provided in Example 2.

2. DNA amplification

The identification of biallelic markers in a sample of genomic DNA may be facilitated through the use of DNA amplification methods. DNA samples can be pooled or unpooled for the amplification step. DNA amplification techniques are well known to those skilled in the art.

Amplification techniques that can be used in the context of the present invention include, but are not limited to, the ligase chain reaction (LCR) described in EP-A- 320 308, WO 9320227 and EP-A-439 182, the polymerase chain reaction (PCR, RT-PCR) and techniques such as the nucleic acid sequence based amplification (NASBA) described in Guatelli J.C., et al.(1990) and in Compton J.(1991), Q-beta amplification as described in European Patent Application No 4544610, strand displacement amplification as described in Walker et al.(1996) and EP A 684 315 and, target mediated amplification as described in PCT Publication WO 9322461.

LCR and Gap LCR are exponential amplification techniques, both depend on DNA ligase to join adjacent primers annealed to a DNA molecule. In Ligase Chain Reaction (LCR), probe pairs are used which include two primary (first and second) and two secondary (third and fourth) probes, all of which are employed in molar excess to target. The first probe hybridizes to a first segment of the target strand and the second probe hybridizes to a second segment of the target strand, the first and second segments being contiguous so that the primary probes abut one another in 5' phosphate-3'hydroxyl relationship, and so that a ligase can covalently fuse or ligate the two probes into a fused product. In addition, a third (secondary) probe can hybridize to a portion of the first probe and a fourth (secondary) probe can hybridize to a portion of the second probe in a similar abutting fashion. Of course, if the target is initially double stranded, the secondary probes also will hybridize to the target complement in the first instance. Once the ligated strand of primary probes is separated from the target strand, it will hybridize with the third and fourth probes, which can be ligated to form a complementary, secondary ligated product. It is important to realize that the ligated products are functionally equivalent to either the target or its complement. By repeated cycles of hybridization and ligation, amplification of the target sequence is achieved. A method for multiplex LCR has also been described (WO 9320227). Gap LCR (GLCR) is a version of LCR where the probes are not adjacent but are separated by 2 to 3 bases.

For amplification of mRNAs, it is within the scope of the present invention to reverse transcribe mRNA into cDNA followed by polymerase chain reaction (RT-PCR); or, to use a single enzyme for both steps as described in U.S. Patent No. 5,322,770 or, to use Asymmetric Gap LCR (RT-AGLCR) as described by Marshall et al.(1994). AGLCR is a modification of GLCR that allows the amplification of RNA.

The PCR technology is the preferred amplification technique used in the present invention. A variety of PCR techniques are familiar to those skilled in the art. For a review of PCR technology, see White (1997) and the publication entitled "PCR Methods and Applications" (1991, Cold Spring Harbor Laboratory Press). In each of these PCR procedures, PCR primers on either side of the nucleic acid sequences to be amplified are added to a suitably prepared nucleic acid sample along with dNTPs and a thermostable polymerase such as Taq polymerase, Pfu polymerase, or Vent polymerase. The nucleic acid in the sample is denatured and the PCR primers are specifically hybridized to complementary nucleic acid sequences in the sample. The hybridized primers are extended. Thereafter, another cycle of denaturation, hybridization, and extension is initiated. The cycles are repeated multiple times to produce an amplified fragment containing the nucleic acid sequence between the primer sites. PCR has further been described in several patents including US Patents 4,683,195; 4,683,202; and 4,965,188.

The PCR technology is the preferred amplification technique used to identify new biallelic markers. A typical example of a PCR reaction suitable for the purposes of the present invention is provided in Example 3.

One of the aspects of the present invention is a method for the amplification of the human *hGGPPS* gene, particularly of a fragment of the genomic sequence of SEQ ID No 1 or of the cDNA sequence of SEQ ID No 2 or 3, or a fragment or a variant thereof in a test sample, preferably using the PCR technology. This method comprises the steps of:

a) contacting a test sample with amplification reaction reagents comprising a pair of amplification primers as described above and located on either side of the polynucleotide region to be amplified, and

b) optionally, detecting the amplification products.

The invention also concerns a kit for the amplification of a *hGGPPS* gene sequence, particularly of a portion of the genomic sequence of SEQ ID No 1 or of the cDNA sequence of SEQ ID No 2 or 3, or a variant thereof in a test sample, wherein said kit comprises:

a) a pair of oligonucleotide primers located on either side of the *hGGPPS* region to be amplified;

b) optionally, the reagents necessary for performing the amplification reaction.

In one embodiment of the above amplification method and kit, the amplification product is detected by hybridization with a labeled probe having a sequence which is complementary to the amplified region. In another embodiment of the above amplification method and kit, primers comprise a sequence which is selected from the group consisting of SEQ ID Nos 7-9.

In a first embodiment of the present invention, biallelic markers are identified using genomic sequence information generated by the inventors. Sequenced genomic DNA fragments are used to design primers for the amplification of 500 bp fragments. These 500 bp fragments are amplified from genomic DNA and are scanned for biallelic markers. Primers may be designed using the OSP software (Hillier L. and Green P., 1991). All primers may contain, upstream of the specific target bases, a common oligonucleotide tail that serves as a sequencing primer. Those skilled in the art are familiar with primer extensions, which can be used for these purposes.

Preferred primers, useful for the amplification of genomic sequences encoding the candidate genes, focus on promoters, exons and splice sites of the genes. A biallelic marker presents a higher probability to be an eventual causal mutation if it is located in these functional regions of the gene. Preferred amplification primers of the invention include the nucleotide sequences of SEQ ID Nos 8 and 9.

Other preferred primers according to the invention allow the amplification of various fragments of the purified or isolated nucleic acid of SEQ ID No 1. These primers are presented below as couples of forward and reverse primers that may be used together to amplify a desired nucleotide sequence.

Position range of forward primers in SEQ ID No 1	Complementary position range of reverse primer in SEQ ID No 1
7233-7251	7565-7582
13582-13600	13982-14001
14222-14240	14626-14645

14606-14623	15007-15026
14845-14864	15246-15265

The primers described above are individually useful as oligonucleotide probes in order to detect the corresponding *hGGPS* nucleotide sequence in a sample, and more preferably to detect the presence of a *hGGPS* DNA molecule in a sample suspected to contain it.

3. Sequencing of amplified genomic DNA and identification of polymorphisms

The amplification products generated as described above, are then sequenced using any method known and available to the skilled technician. Methods for sequencing DNA using either the dideoxy-mediated method (Sanger method) or the Maxam-Gilbert method are widely known to those of ordinary skill in the art. Such methods are for example disclosed in Sambrook et al.(1989). Alternative approaches include hybridization to high-density DNA probe arrays as described in Chee et al.(1996).

Preferably, the amplified DNA is subjected to automated dideoxy terminator sequencing reactions using a dye-primer cycle sequencing protocol. The products of the sequencing reactions are run on sequencing gels and the sequences are determined using gel image analysis. The polymorphism search is based on the presence of superimposed peaks in the electrophoresis pattern resulting from different bases occurring at the same position. Because each dideoxy terminator is labeled with a different fluorescent molecule, the two peaks corresponding to a biallelic site present distinct colors corresponding to two different nucleotides at the same position on the sequence. However, the presence of two peaks can be an artifact due to background noise. To exclude such an artifact, the two DNA strands are sequenced and a comparison between the peaks is carried out. In order to be registered as a polymorphic sequence, the polymorphism has to be detected on both strands.

The above procedure permits those amplification products, which contain biallelic markers to be identified. The detection limit for the frequency of biallelic polymorphisms detected by sequencing pools of 100 individuals is approximately 0.1 for the minor allele, as verified by sequencing pools of known allelic frequencies. However, more than 90% of the biallelic polymorphisms detected by the pooling method have a frequency for the minor allele higher than 0.25. Therefore, the biallelic markers selected by this method have a frequency of at least 0.1 for the minor allele and less than 0.9 for the major allele. Preferably at least 0.2 for the minor allele and less than 0.8 for the major allele, more preferably at least 0.3 for the minor allele and less than 0.7 for the major allele, thus a heterozygosity rate higher than 0.18, preferably higher than 0.32, more preferably higher than 0.42.

In another embodiment, biallelic markers are detected by sequencing individual DNA samples, the frequency of the minor allele of such a biallelic marker may be less than 0.1.

In a particular embodiment of the invention, the test samples are a pool of 100 individuals and 50 individual samples. This is the methodology used in the preferred embodiment of the present

invention, in which 1 biallelic marker has been identified in a genomic region containing the *hGGPS* gene. This biallelic marker is called 5-187-77 and is located in intron 3 of *hGGPPS* gene. The biallelic marker consists in an insertion of a nucleotide T.

The polymorphisms identified above can be further confirmed and their respective frequencies can be determined through various methods using the previously described primers and probes as described herein. These methods can also be useful for genotyping either new populations in association studies or linkage analysis or individuals in the context of detection of alleles of biallelic markers which are known to be associated with a given trait. The genotyping of the biallelic markers is also important for the mapping. It will be appreciated that the methods described below can be equally performed on individual or pooled DNA samples.

b) Genotyping Of Biallelic Markers

Once a given polymorphic site has been found and characterized as a biallelic marker as described above, several methods can be used in order to determine the specific allele carried by an individual at the given polymorphic base.

The identification of biallelic markers described previously allows the design of appropriate oligonucleotides, which can be used as probes and primers, to amplify a *hGGPS* gene containing the polymorphic site of interest and for the detection of such polymorphisms.

In one embodiment the invention encompasses methods of genotyping comprising determining the identity of a nucleotide at a *hGGPPS*-related biallelic marker or the complement thereof in a biological sample; optionally, wherein said *hGGPPS* -related biallelic marker is the biallelic marker 5-187-77, and the complement thereof; optionally, wherein said biological sample is derived from a single subject; optionally, wherein the identity of the nucleotides at said biallelic marker is determined for both copies of said biallelic marker present in said individual's genome; optionally, wherein said biological sample is derived from multiple subjects; Optionally, the genotyping methods of the invention encompass methods with any further limitation described in this disclosure, or those following, specified alone or in any combination; Optionally, said method is performed *in vitro*; optionally, further comprising amplifying a portion of said sequence comprising the biallelic marker prior to said determining step; Optionally, wherein said amplifying is performed by PCR, LCR, or replication of a recombinant vector comprising an origin of replication and said fragment in a host cell; optionally, wherein said determining is performed by a hybridization assay, a sequencing assay, a microsequencing assay, or an enzyme-based mismatch detection assay.

1) Amplification

Methods and polynucleotides are provided to amplify a segment of nucleotides comprising one or more biallelic marker of the present invention. It will be appreciated that amplification of DNA fragments comprising biallelic markers may be used in various methods and for various purposes and is not restricted t genotyping. Nevertheless, many genotyping methods, although not

all, require the previous amplification of the DNA region carrying the biallelic marker of interest. Such methods specifically increase the concentration or total number of sequences that span the biallelic marker or include that site and sequences located either distal or proximal to it. Diagnostic assays may also rely on amplification of DNA segments carrying a biallelic marker of the present invention. Amplification of DNA may be achieved by any method known in the art. Amplification techniques are described above in the section entitled, "DNA amplification."

Some of these amplification methods are particularly suited for the detection of single nucleotide polymorphisms and allow the simultaneous amplification of a target sequence and the identification of the polymorphic nucleotide as it is further described below.

The identification of biallelic markers as described above allows the design of appropriate oligonucleotides, which can be used as primers to amplify DNA fragments comprising the biallelic markers of the present invention. Amplification can be performed using the primers initially used to discover new biallelic markers which are described herein or any set of primers allowing the amplification of a DNA fragment comprising a biallelic marker of the present invention.

In some embodiments the present invention provides primers for amplifying a DNA fragment containing one or more biallelic markers of the present invention. Preferred amplification primers are listed in Example 3. It will be appreciated that the primers listed are merely exemplary and that any other set of primers which produce amplification products containing one or more biallelic markers of the present invention are also of use.

The spacing of the primers determines the length of the segment to be amplified. In the context of the present invention, amplified segments carrying biallelic markers can range in size from at least about 25 bp to 35 kbp. Amplification fragments from 25-3000 bp are typical, fragments from 50-1000 bp are preferred and fragments from 100-600 bp are highly preferred. It will be appreciated that amplification primers for the biallelic markers may be any sequence which allow the specific amplification of any DNA fragment carrying the markers. Amplification primers may be labeled or immobilized on a solid support as described in "Oligonucleotide probes and primers".

2) Sequencing

The nucleotide present at a polymorphic site can be determined by sequencing methods. In a preferred embodiment, DNA samples are subjected to PCR amplification before sequencing as described above. DNA sequencing methods are described in "Sequencing Of Amplified Genomic DNA And Identification Of Polymorphisms".

Preferably, the amplified DNA is subjected to automated dideoxy terminator sequencing reactions using a dye-primer cycle sequencing protocol. Sequence analysis allows the identification of the base present at the biallelic marker site.

3) Microsequencing

In microsequencing methods, the nucleotide at a polymorphic site in a target DNA is detected by a single nucleotide primer extension reaction. This method involves appropriate microsequencing primers which, hybridize just upstream of the polymorphic base of interest in the target nucleic acid. A polymerase is used to specifically extend the 3' end of the primer with one single ddNTP (chain terminator) complementary to the nucleotide at the polymorphic site. Next the identity of the incorporated nucleotide is determined in any suitable way.

Typically, microsequencing reactions are carried out using fluorescent ddNTPs and the extended microsequencing primers are analyzed by electrophoresis on ABI 377 sequencing machines to determine the identity of the incorporated nucleotide as described in EP 412 883, the disclosure of which is incorporated herein by reference in its entirety. Alternatively capillary electrophoresis can be used in order to process a higher number of assays simultaneously. An example of a typical microsequencing procedure that can be used in the context of the present invention is provided in Example 5.

Different approaches can be used for the labeling and detection of ddNTPs. A homogeneous phase detection method based on fluorescence resonance energy transfer has been described by Chen and Kwok (1997) and Chen et al.(1997). Alternatively, the extended primer may be analyzed by MALDI-TOF Mass Spectrometry. The base at the polymorphic site is identified by the mass added onto the microsequencing primer (see Haff and Smirnov, 1997).

Microsequencing may be achieved by the established microsequencing method or by developments or derivatives thereof. Alternative methods include several solid-phase microsequencing techniques. The basic microsequencing protocol is the same as described previously, except that the method is conducted as a heterogeneous phase assay, in which the primer or the target molecule is immobilized or captured onto a solid support. For example, immobilization can be carried out via an interaction between biotinylated DNA and streptavidin-coated microtitration wells or avidin-coated polystyrene particles. In the same manner, oligonucleotides or templates may be attached to a solid support in a high-density format. In such solid phase microsequencing reactions, incorporated ddNTPs can be radiolabeled (Syvänen, 1994) or linked to fluorescein (Livak and Hainer, 1994). The detection of radiolabeled ddNTPs can be achieved through scintillation-based techniques. The detection of fluorescein-linked ddNTPs can be based on the binding of antifluorescein antibody conjugated with alkaline phosphatase, followed by incubation with a chromogenic substrate (such as *p*-nitrophenyl phosphate). Other possible reporter-detection pairs include: ddNTP linked to dinitrophenyl (DNP) and anti-DNP alkaline phosphatase conjugate (Harju et al., 1993) or biotinylated ddNTP and horseradish peroxidase-conjugated streptavidin with *o*-phenylenediamine as a substrate (WO 92/15712). As yet another alternative solid-phase microsequencing procedure, Nyren et al.(1993) described a method relying on the

detection of DNA polymerase activity by an enzymatic luminometric inorganic pyrophosphate detection assay (ELIDA).

Pastinen et al.(1997) describe a method for multiplex detection of single nucleotide polymorphism in which the solid phase minisequencing principle is applied to an oligonucleotide array format. High-density arrays of DNA probes attached to a solid support (DNA chips) are further described below.

In one aspect the present invention provides polynucleotides and methods to genotype one or more biallelic markers of the present invention by performing a microsequencing assay. Preferred microsequencing primers include the nucleotide sequence of SEQ ID No 7. It will be appreciated that the microsequencing primer of SEQ ID No 7 is merely exemplary and that, any primer having a 3' end immediately adjacent to the polymorphic nucleotide may be used. Similarly, it will be appreciated that microsequencing analysis may be performed for any biallelic marker or any combination of biallelic markers of the present invention. One aspect of the present invention is a solid support which includes one or more microsequencing primers for determining the identity of a nucleotide at a biallelic marker site.

4. Mismatch detection assays based on polymerases and ligases

In one aspect the present invention provides polynucleotides and methods to determine the allele of one or more biallelic markers of the present invention in a biological sample, by mismatch detection assays based on polymerases and/or ligases. These assays are based on the specificity of polymerases and ligases. Polymerization reactions places particularly stringent requirements on correct base pairing of the 3' end of the amplification primer and the joining of two oligonucleotides hybridized to a target DNA sequence is quite sensitive to mismatches close to the ligation site, especially at the 3' end. Methods, primers and various parameters to amplify DNA fragments comprising biallelic markers of the present invention are further described above in "DNA amplification".

Allele Specific Amplification Primers

Discrimination between the two alleles of a biallelic marker can also be achieved by allele specific amplification, a selective strategy, whereby one of the alleles is amplified without amplification of the other allele. For allele specific amplification, at least one member of the pair of primers is sufficiently complementary with a region of a *hGGPPS* gene comprising the polymorphic base of a biallelic marker of the present invention to hybridize therewith and to initiate the amplification. Such primers are able to discriminate between the two alleles of a biallelic marker.

This is accomplished by placing the polymorphic base at the 3' end of one of the amplification primers. Because the extension forms from the 3'end of the primer, a mismatch at or near this position has an inhibitory effect on amplification. Therefore, under appropriate amplification conditions, these primers only direct amplification on their complementary allele.

Determining the precise location of the mismatch and the corresponding assay conditions are well within the ordinary skill in the art.

Ligation/Amplification Based Methods

The "Oligonucleotide Ligation Assay" (OLA) uses two oligonucleotides which are designed to be capable of hybridizing to abutting sequences of a single strand of a target molecules. One of the oligonucleotides is biotinylated, and the other is detectably labeled. If the precise complementary sequence is found in a target molecule, the oligonucleotides will hybridize such that their termini abut, and create a ligation substrate that can be captured and detected. OLA is capable of detecting single nucleotide polymorphisms and may be advantageously combined with PCR as described by Nickerson et al.(1990). In this method, PCR is used to achieve the exponential amplification of target DNA, which is then detected using OLA.

Other amplification methods which are particularly suited for the detection of single nucleotide polymorphism include LCR (ligase chain reaction), Gap LCR (GLCR) which are described above in "DNA Amplification". LCR uses two pairs of probes to exponentially amplify a specific target. The sequences of each pair of oligonucleotides, is selected to permit the pair to hybridize to abutting sequences of the same strand of the target. Such hybridization forms a substrate for a template-dependant ligase. In accordance with the present invention, LCR can be performed with oligonucleotides having the proximal and distal sequences of the same strand of a biallelic marker site. In one embodiment, either oligonucleotide will be designed to include the biallelic marker site. In such an embodiment, the reaction conditions are selected such that the oligonucleotides can be ligated together only if the target molecule either contains or lacks the specific nucleotide that is complementary to the biallelic marker on the oligonucleotide. In an alternative embodiment, the oligonucleotides will not include the biallelic marker, such that when they hybridize to the target molecule, a "gap" is created as described in WO 90/01069. This gap is then "filled" with complementary dNTPs (as mediated by DNA polymerase), or by an additional pair of oligonucleotides. Thus at the end of each cycle, each single strand has a complement capable of serving as a target during the next cycle and exponential allele-specific amplification of the desired sequence is obtained.

Ligase/Polymerase-mediated Genetic Bit Analysis™ is another method for determining the identity of a nucleotide at a preselected site in a nucleic acid molecule (WO 95/21271). This method involves the incorporation of a nucleoside triphosphate that is complementary to the nucleotide present at the preselected site onto the terminus of a primer molecule, and their subsequent ligation to a second oligonucleotide. The reaction is monitored by detecting a specific label attached to the reaction's solid phase or by detection in solution.

5. Hybridization Assay Methods

A preferred method of determining the identity of the nucleotide present at a biallelic marker site involves nucleic acid hybridization. The hybridization probes, which can be conveniently used

in such reactions, preferably include the probes defined herein. Any hybridization assay may be used including Southern hybridization, Northern hybridization, dot blot hybridization and solid-phase hybridization (see Sambrook et al., 1989).

Specific probes can be designed that hybridize to one form of a biallelic marker and not to the other and therefore are able to discriminate between different allelic forms. Allele-specific probes are often used in pairs, one member of a pair showing perfect match to a target sequence containing the original allele and the other showing a perfect match to the target sequence containing the alternative allele. Hybridization conditions should be sufficiently stringent that there is a significant difference in hybridization intensity between alleles, and preferably an essentially binary response, whereby a probe hybridizes to only one of the alleles. Stringent, sequence specific hybridization conditions, under which a probe will hybridize only to the exactly complementary target sequence are well known in the art (Sambrook et al., 1989). Although such hybridization can be performed in solution, it is preferred to employ a solid-phase hybridization assay. The target DNA comprising a biallelic marker of the present invention may be amplified prior to the hybridization reaction. The presence of a specific allele in the sample is determined by detecting the presence or the absence of stable hybrid duplexes formed between the probe and the target DNA. The detection of hybrid duplexes can be carried out by a number of methods. Various detection assay formats are well known which utilize detectable labels bound to either the target or the probe to enable detection of the hybrid duplexes. Typically, hybridization duplexes are separated from unhybridized nucleic acids and the labels bound to the duplexes are then detected. Those skilled in the art will recognize that wash steps may be employed to wash away excess target DNA or probe as well as unbound conjugate. Further, standard heterogeneous assay formats are suitable for detecting the hybrids using the labels present on the primers and probes.

Two recently developed assays allow hybridization-based allele discrimination with no need for separations or washes (see Landegren U. et al., 1998). The TaqMan assay takes advantage of the 5' nuclease activity of Taq DNA polymerase to digest a DNA probe annealed specifically to the accumulating amplification product. TaqMan probes are labeled with a donor-acceptor dye pair that interacts via fluorescence energy transfer. Cleavage of the TaqMan probe by the advancing polymerase during amplification dissociates the donor dye from the quenching acceptor dye, greatly increasing the donor fluorescence. All reagents necessary to detect two allelic variants can be assembled at the beginning of the reaction and the results are monitored in real time (see Livak et al., 1995). In an alternative homogeneous hybridization based procedure, molecular beacons are used for allele discriminations. Molecular beacons are hairpin-shaped oligonucleotide probes that report the presence of specific nucleic acids in homogeneous solutions. When they bind to their targets they undergo a conformational reorganization that restores the fluorescence of an internally quenched fluorophore (Tyagi et al., 1998).

The polynucleotides provided herein can be used to produce probes which can be used in hybridization assays for the detection of biallelic marker alleles in biological samples. These probes are characterized in that they preferably comprise between 8 and 50 nucleotides, and in that they are sufficiently complementary to a sequence comprising a biallelic marker of the present invention to hybridize thereto and preferably sufficiently specific to be able to discriminate the targeted sequence for only one nucleotide variation. A particularly preferred probe is 25 nucleotides in length. Preferably the biallelic marker is within 4 nucleotides of the center of the polynucleotide probe. In particularly preferred probes, the biallelic marker is at the center of said polynucleotide. Preferred probes comprise a nucleotide sequence selected from the group consisting of amplicons listed in Table 1 and the sequences complementary thereto, or a fragment thereof, said fragment comprising at least about 8 consecutive nucleotides, preferably 10, 15, 20, more preferably 25, 30, 40, 47, or 50 consecutive nucleotides and containing a polymorphic base. Preferred probes comprise a nucleotide sequence selected from the group consisting of SEQ ID Nos 5 and 6 and the sequences complementary thereto. In preferred embodiments the polymorphic base(s) are within 5, 4, 3, 2, 1, nucleotides of the center of the said polynucleotide, more preferably at the center of said polynucleotide.

Preferably the probes of the present invention are labeled or immobilized on a solid support. Labels and solid supports are further described in "Oligonucleotide Probes and Primers". The probes can be non-extendable as described in "Oligonucleotide Probes and Primers".

By assaying the hybridization to an allele specific probe, one can detect the presence or absence of a biallelic marker allele in a given sample. High-Throughput parallel hybridization in array format is specifically encompassed within "hybridization assays" and are described below.

6- Hybridization To Addressable Arrays Of Oligonucleotides

Hybridization assays based on oligonucleotide arrays rely on the differences in hybridization stability of short oligonucleotides to perfectly matched and mismatched target sequence variants. Efficient access to polymorphism information is obtained through a basic structure comprising high-density arrays of oligonucleotide probes attached to a solid support (e.g., the chip) at selected positions. Each DNA chip can contain thousands to millions of individual synthetic DNA probes arranged in a grid-like pattern and miniaturized to the size of a dime.

The chip technology has already been applied with success in numerous cases. For example, the screening of mutations has been undertaken in the BRCA1 gene, in *S. cerevisiae* mutant strains, and in the protease gene of HIV-1 virus (Hacia et al., 1996; Shoemaker et al., 1996; Kozal et al., 1996). Chips of various formats for use in detecting biallelic polymorphisms can be produced on a customized basis by Affymetrix (GeneChip™), Hyseq (HyChip and HyGnostics), and Protogene Laboratories.

In general, these methods employ arrays of oligonucleotide probes that are complementary to target nucleic acid sequence segments from an individual which, target sequences include a

polymorphic marker. EP 785280 describes a tiling strategy for the detection of single nucleotide polymorphisms. Briefly, arrays may generally be "tiled" for a large number of specific polymorphisms. By "tiling" is generally meant the synthesis of a defined set of oligonucleotide probes which is made up of a sequence complementary to the target sequence of interest, as well as preselected variations of that sequence, e.g., substitution of one or more given positions with one or more members of the basis set of nucleotides. Tiling strategies are further described in PCT application No. WO 95/11995. Hybridization and scanning may be carried out as described in PCT application No. WO 92/10092 and WO 95/11995 and US patent No. 5,424,186.

Thus, in some embodiments, the chips may comprise an array of nucleic acid sequences of fragments of about 15 nucleotides in length. In further embodiments, the chip may comprise an array including at least one of the sequences selected from the group consisting of amplicons listed in table 1 and the sequences complementary thereto, or a fragment thereof, said fragment comprising at least about 8 consecutive nucleotides, preferably 10, 15, 20, more preferably 25, 30, 40, 47, or 50 consecutive nucleotides and containing a polymorphic base. In preferred embodiments the polymorphic base is within 5, 4, 3, 2, 1, nucleotides of the center of the said polynucleotide, more preferably at the center of said polynucleotide. In some embodiments, the chip may comprise an array of at least 2, 3, 4, 5, 6, 7, 8 or more of these polynucleotides of the invention. Solid supports and polynucleotides of the present invention attached to solid supports are further described in "Oligonucleotide Probes And Primers".

7- Integrated Systems

Another technique, which may be used to analyze polymorphisms, includes multicomponent integrated systems, which miniaturize and compartmentalize processes such as PCR and capillary electrophoresis reactions in a single functional device. An example of such technique is disclosed in US patent 5,589,136, which describes the integration of PCR amplification and capillary electrophoresis in chips.

Integrated systems can be envisaged mainly when microfluidic systems are used. These systems comprise a pattern of microchannels designed onto a glass, silicon, quartz, or plastic wafer included on a microchip. The movements of the samples are controlled by electric, electroosmotic or hydrostatic forces applied across different areas of the microchip to create functional microscopic valves and pumps with no moving parts.

For genotyping biallelic markers, the microfluidic system may integrate nucleic acid amplification, microsequencing, capillary electrophoresis and a detection method such as laser-induced fluorescence detection.

Oligonucleotide Probes and primers

Polynucleotides derived from the *hGGPPS* gene are useful in order to detect the presence of at least a copy of a nucleotide sequence of SEQ ID No 1, or a fragment, complement, or variant

thereof in a test sample. Furthermore polynucleotides derived from the *hGGPPS* gene can be used to generate antisense polynucleotide or polynucleotide for the triple helix strategy.

Particularly preferred probes and primers of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of SEQ ID No 1 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of the following nucleotide positions of SEQ ID No 1: 1-485, 547-632, 827-7291, 7385-13759, 13831-14062, 14671-15054, and 15252-17131.

The invention also relates to nucleic acid probes characterized in that they hybridize specifically, under the stringent hybridization conditions defined above, with a nucleic acid selected from the group consisting of the nucleotide sequences 1-485, 547-632, 827-7291, 7385-13759, 13831-14062, 14671-15054, and 15252-17131 of SEQ ID No 1 or a variant thereof or a sequence complementary thereto.

Particularly preferred probes and primers of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of SEQ ID No 2 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of the nucleotide positions 834-1217 of SEQ ID No 2. Additional preferred probes and primers of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, or 1000 nucleotides of SEQ ID No 2 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 of the nucleotide positions 967-1351 of SEQ ID No 3.

The invention also relates to nucleic acid probes characterized in that they hybridize specifically, under the stringent hybridization conditions defined above, with a nucleic acid selected from the group consisting of the nucleotide sequences 834-1217 of SEQ ID No 2 and 967-1351 of SEQ ID No 3, or a variant thereof or a sequence complementary thereto.

In one embodiment the invention encompasses isolated, purified, and recombinant polynucleotides consisting of, or consisting essentially of a contiguous span of 8 to 50 nucleotides of any one of SEQ ID Nos 1-3 and the complement thereof, wherein said span includes a *hGGPPS*-related biallelic marker in said sequence; optionally, wherein said *hGGPPS*-related biallelic marker is the biallelic marker 5-187-77, and the complement thereof; optionally, wherein said contiguous span is 18 to 50 nucleotides in length and said biallelic marker is within 4 nucleotides of the center of said polynucleotide; optionally, wherein said polynucleotide consists of said contiguous span and said contiguous span is 25 nucleotides in length and said biallelic marker is at the center of said polynucleotide; optionally, wherein the 3' end of said contiguous span is present at the 3' end of said polynucleotide; and optionally, wherein the 3' end of said contiguous span is located at the 3' end of said polynucleotide and said biallelic marker is present at the 3' end of said polynucleotide. In a

preferred embodiment, said probes comprises, consists of, or consists essentially of a sequence selected from SEQ ID Nos 5 and 6 and the complementary sequences thereto.

In another embodiment the invention encompasses isolated, purified and recombinant polynucleotides comprising, consisting of, or consisting essentially of a contiguous span of 8 to 50 nucleotides of SEQ ID Nos 1-3, or the complements thereof, wherein the 3' end of said contiguous span is located at the 3' end of said polynucleotide, and wherein the 3' end of said polynucleotide is located within 20 nucleotides upstream of a *hGGPPS* -related biallelic marker in said sequence; optionally, wherein said *hGGPPS*-related biallelic marker is the biallelic marker 5-187-77, and the complement thereof; optionally, wherein the 3' end of said polynucleotide is located 1 nucleotide upstream of said *hGGPPS* -related biallelic marker in said sequence; and optionally, wherein said polynucleotide consists essentially of a sequence of SEQ ID No 7.

In a further embodiment, the invention encompasses isolated, purified, or recombinant polynucleotides comprising, consisting of, or consisting essentially of a sequence selected from the sequences of SEQ ID Nos 8and 9.

In an additional embodiment, the invention encompasses polynucleotides for use in hybridization assays, sequencing assays, and enzyme-based mismatch detection assays for determining the identity of the nucleotide at a *hGGPPS* -related biallelic marker, as well as polynucleotides for use in amplifying segments of nucleotides comprising a *hGGPPS* -related biallelic marker; optionally, wherein said *hGGPPS*-related biallelic marker is the biallelic marker 5-187-77, and the complements thereof.

A probe or a primer according to the invention has between 8 and 1000 nucleotides in length, or is specified to be at least 12, 15, 18, 20, 25, 35, 40, 50, 60, 70, 80, 100, 250, 500 or 1000 nucleotides in length. More particularly, the length of these probes and primers can range from 8, 10, 15, 20, or 30 to 100 nucleotides, preferably from 10 to 50, more preferably from 15 to 30 nucleotides. The appropriate length for primers and probes under a particular set of assay conditions may be empirically determined by one of skill in the art. A preferred probe or primer consists of a nucleic acid comprising a polynucleotide selected from the group of the nucleotide sequences of SEQ ID Nos 5-9 or a fragment thereof or a complementary sequence thereto.

The formation of stable hybrids depends on the melting temperature (Tm) of the DNA. The Tm depends on the length of the primer or probe, the ionic strength of the solution and the G+C content. The higher the G+C content of the primer or probe, the higher is the melting temperature because G:C pairs are held by three H bonds whereas A:T pairs have only two. The GC content in the probes of the invention usually ranges between 10 and 75 %, preferably between 35 and 60 %, and more preferably between 40 and 55 %.

The primers and probes can be prepared by any suitable method, including, for example, cloning and restriction of appropriate sequences and direct chemical synthesis by a method such as the phosphodiester method of Narang et al.(1979), the phosphodiester method of Brown et al.(1979),

the diethylphosphoramidite method of Beaucage et al.(1981) and the solid support method described in EP 0 707 592.

Detection probes are generally nucleic acid sequences or uncharged nucleic acid analogs such as, for example peptide nucleic acids which are disclosed in International Patent Application WO 92/20702, morpholino analogs which are described in U.S. Patents Numbered 5,185,444; 5,034,506 and 5,142,047. The probe may have to be rendered "non-extendable" in that additional dNTPs cannot be added to the probe. In and of themselves analogs usually are non-extendable and nucleic acid probes can be rendered non-extendable by modifying the 3' end of the probe such that the hydroxyl group is no longer capable of participating in elongation. For example, the 3' end of the probe can be functionalized with the capture or detection label to thereby consume or otherwise block the hydroxyl group. Alternatively, the 3' hydroxyl group simply can be cleaved, replaced or modified, U.S. Patent Application Serial No. 07/049,061 filed April 19, 1993 describes modifications, which can be used to render a probe non-extendable.

Any of the polynucleotides of the present invention can be labeled, if desired, by incorporating any label known in the art to be detectable by spectroscopic, photochemical, biochemical, immunochemical, or chemical means. For example, useful labels include radioactive substances (including, ^{32}P, ^{35}S, ^{3}H, ^{125}I), fluorescent dyes (including, 5-bromodesoxyuridin, fluorescein, acetylaminofluorene, digoxigenin) or biotin. Preferably, polynucleotides are labeled at their 3' and 5' ends. Examples of non-radioactive labeling of nucleic acid fragments are described in the French patent No. FR-7810975 or by Urdea et al (1988) or Sanchez-Pescador et al (1988). In addition, the probes according to the present invention may have structural characteristics such that they allow the signal amplification, such structural characteristics being, for example, branched DNA probes as those described by Urdea et al. in 1991 or in the European patent No. EP 0 225 807 (Chiron).

A label can also be used to capture the primer, so as to facilitate the immobilization of either the primer or a primer extension product, such as amplified DNA, on a solid support. A capture label is attached to the primers or probes and can be a specific binding member which forms a binding pair with the solid's phase reagent's specific binding member (e.g. biotin and streptavidin). Therefore depending upon the type of label carried by a polynucleotide or a probe, it may be employed to capture or to detect the target DNA. Further, it will be understood that the polynucleotides, primers or probes provided herein, may, themselves, serve as the capture label. For example, in the case where a solid phase reagent's binding member is a nucleic acid sequence, it may be selected such that it binds a complementary portion of a primer or probe to thereby immobilize the primer or probe to the solid phase. In cases where a polynucleotide probe itself serves as the binding member, those skilled in the art will recognize that the probe will contain a sequence or "tail" that is not complementary to the target. In the case where a polynucleotide primer

itself serves as the capture label, at least a portion of the primer will be free to hybridize with a nucleic acid on a solid phase. DNA Labeling techniques are well known to the skilled technician.

The probes of the present invention are useful for a number of purposes. They can be notably used in Southern hybridization to genomic DNA. The probes can also be used to detect PCR amplification products. They may also be used to detect mismatches in the *hGGPPS* gene or mRNA using other techniques.

Any of the polynucleotides, primers and probes of the present invention can be conveniently immobilized on a solid support. Solid supports are known to those skilled in the art and include the walls of wells of a reaction tray, test tubes, polystyrene beads, magnetic beads, nitrocellulose strips, membranes, microparticles such as latex particles, sheep (or other animal) red blood cells, duracytes and others. The solid support is not critical and can be selected by one skilled in the art. Thus, latex particles, microparticles, magnetic or non-magnetic beads, membranes, plastic tubes, walls of microtiter wells, glass or silicon chips, sheep (or other suitable animal's) red blood cells and duracytes are all suitable examples. Suitable methods for immobilizing nucleic acids on solid phases include ionic, hydrophobic, covalent interactions and the like. A solid support, as used herein, refers to any material which is insoluble, or can be made insoluble by a subsequent reaction. The solid support can be chosen for its intrinsic ability to attract and immobilize the capture reagent. Alternatively, the solid phase can retain an additional receptor which has the ability to attract and immobilize the capture reagent. The additional receptor can include a charged substance that is oppositely charged with respect to the capture reagent itself or to a charged substance conjugated to the capture reagent. As yet another alternative, the receptor molecule can be any specific binding member which is immobilized upon (attached to) the solid support and which has the ability to immobilize the capture reagent through a specific binding reaction. The receptor molecule enables the indirect binding of the capture reagent to a solid support material before the performance of the assay or during the performance of the assay. The solid phase thus can be a plastic, derivatized plastic, magnetic or non-magnetic metal, glass or silicon surface of a test tube, microtiter well, sheet, bead, microparticle, chip, sheep (or other suitable animal's) red blood cells, duracytes® and other configurations known to those of ordinary skill in the art. The polynucleotides of the invention can be attached to or immobilized on a solid support individually or in groups of at least 2, 5, 8, 10, 12, 15, 20, or 25 distinct polynucleotides of the invention to a single solid support. In addition, polynucleotides other than those of the invention may be attached to the same solid support as one or more polynucleotides of the invention.

Consequently, the invention also deals with a method for detecting the presence of a nucleic acid comprising at least a part of a nucleotide sequence selected from the group consisting of SEQ ID Nos 1-3 in a sample, said method comprising the following steps of :

a) bringing into contact a nucleic acid probe or a plurality of nucleic acid probes, which can hybridize to a nucleotide sequence included in one of the nucleic acids of SEQ ID Nos 1-3, and the sample to be assayed.

b) detecting the hybrid complex formed between the probe and a nucleic acid in the sample.

Preferably, the nucleic acid probe is selected from the group of polynucleotides consisting of the nucleotide sequences SEQ ID Nos 5-9. In a first preferred embodiment of this detection method, said nucleic acid probe or the plurality of nucleic acid probes are labeled with a detectable molecule. In a second preferred embodiment of said method, said nucleic acid probe or the plurality of nucleic acid probes has been immobilized on a substrate.

The invention further concerns a kit for detecting the presence of a nucleic acid comprising at least a part of a nucleotide sequence selected from the group consisting of SEQ ID Nos 1-3 in a sample, said kit comprising :

a) a nucleic acid probe or a plurality of nucleic acid probes which can hybridize to a nucleotide sequence included in one of the nucleic acids of SEQ ID Nos 1-3;

b) optionally, the reagents necessary for performing the hybridization reaction.

The nucleic acid probe or the plurality of nucleic acid probes that are included in the detection kit described above may be selected from the group consisting of SEQ ID Nos 5-9. In a first preferred embodiment of the detection kit, the nucleic acid probe or the plurality of nucleic acid probes are labeled with a detectable molecule. In a second preferred embodiment of the detection kit, the nucleic acid probe or the plurality of nucleic acid probes has been immobilized on a substrate.

Oligonucleotide arrays

A substrate comprising a plurality of oligonucleotide primers or probes of the invention may be used either for detecting or amplifying targeted sequences in the *hGGPPS* gene and may also be used for detecting mutations in the coding or in the non-coding sequences of the *hGGPPS* gene.

Any polynucleotide provided herein may be attached in overlapping areas or at random locations on the solid support. Alternatively the polynucleotides of the invention may be attached in an ordered array wherein each polynucleotide is attached to a distinct region of the solid support which does not overlap with the attachment site of any other polynucleotide. Preferably, such an ordered array of polynucleotides is designed to be "addressable" where the distinct locations are recorded and can be accessed as part of an assay procedure. Addressable polynucleotide arrays typically comprise a plurality of different oligonucleotide probes that are coupled to a surface of a substrate in different known locations. The knowledge of the precise location of each polynucleotides location makes these "addressable" arrays particularly useful in hybridization assays. Any addressable array technology known in the art can be employed with the polynucleotides of the invention. One particular embodiment of these polynucleotide arrays is known as the Genechips™, and has been generally described in US Patent 5,143,854; PCT publications WO 90/15070 and 92/10092. These arrays may generally be produced using

mechanical synthesis methods or light directed synthesis methods which incorporate a combination of photolithographic methods and solid phase oligonucleotide synthesis (Fodor et al., 1991). The immobilization of arrays of oligonucleotides on solid supports has been rendered possible by the development of a technology generally identified as "Very Large Scale Immobilized Polymer Synthesis" (VLSIPS™) in which, typically, probes are immobilized in a high density array on a solid surface of a chip. Examples of VLSIPS™ technologies are provided in US Patents 5,143,854; and 5,412,087 and in PCT Publications WO 90/15070, WO 92/10092 and WO 95/11995, which describe methods for forming oligonucleotide arrays through techniques such as light-directed synthesis techniques. In designing strategies aimed at providing arrays of nucleotides immobilized on solid supports, further presentation strategies were developed to order and display the oligonucleotide arrays on the chips in an attempt to maximize hybridization patterns and sequence information. Examples of such presentation strategies are disclosed in PCT Publications WO 94/12305, WO 94/11530, WO 97/29212 and WO 97/31256.

In another embodiment of the oligonucleotide arrays of the invention, an oligonucleotide probe matrix may advantageously be used to detect mutations occurring in the *hGGPPS* gene and preferably in its regulatory region. For this particular purpose, probes are specifically designed to have a nucleotide sequence allowing their hybridization to the genes that carry known mutations (either by deletion, insertion or substitution of one or several nucleotides). By known mutations, it is meant, mutations on the *hGGPPS* gene that have been identified according, for example to the technique used by Huang et al.(1996) or Samson et al.(1996).

Another technique that is used to detect mutations in the *hGGPPS* gene is the use of a high-density DNA array. Each oligonucleotide probe constituting a unit element of the high density DNA array is designed to match a specific subsequence of the *hGGPPS* genomic DNA or cDNA. Thus, an array consisting of oligonucleotides complementary to subsequences of the target gene sequence is used to determine the identity of the target sequence with the wild gene sequence, measure its amount, and detect differences between the target sequence and the reference wild gene sequence of the *hGGPPS* gene. In one such design, termed 4L tiled array, is implemented a set of four probes (A, C, G, T), preferably 15-nucleotide oligomers. In each set of four probes, the perfect complement will hybridize more strongly than mismatched probes. Consequently, a nucleic acid target of length L is scanned for mutations with a tiled array containing 4L probes, the whole probe set containing all the possible mutations in the known wild reference sequence. The hybridization signals of the 15-mer probe set tiled array are perturbed by a single base change in the target sequence. As a consequence, there is a characteristic loss of signal or a "footprint" for the probes flanking a mutation position. This technique was described by Chee et al. in 1996.

Consequently, the invention concerns an array of nucleic acid molecules comprising at least one polynucleotide described above as probes and primers. Preferably, the invention concerns an array of nucleic acid comprising at least two polynucleotides described above as probes and primers.

A further object of the invention consists of an array of nucleic acid sequences comprising either at least one of the sequences selected from the group consisting of SEQ ID Nos 5-9, the sequences complementary thereto, a fragment thereof of at least 8, 10, 12, 15, 18, 20, 25, 30, or 40 consecutive nucleotides thereof, and at least one sequence comprising the biallelic marker 5-187-77 and the complements thereto.

The invention also pertains to an array of nucleic acid sequences comprising either at least two of the sequences selected from the group consisting of SEQ ID Nos 5-9, the sequences complementary thereto, a fragment thereof of at least 8, 10, 12, 15, 18, 20, 25, 30, or 40 consecutive nucleotides thereof, and at least one sequence comprising the biallelic marker 5-187-77 and the complements thereto.

Vectors for the expression of a regulatory or a coding polynucleotide according to the invention.

Any of the regulatory polynucleotides or the coding polynucleotides of the invention may be inserted into recombinant vectors for expression in a recombinant host cell or a recombinant host organism.

Thus, the present invention also encompasses a family of recombinant vectors that contains either a regulatory polynucleotide selected from the group consisting of the regulatory polynucleotides derived from the *hGGPS* gene, or a polynucleotide comprising the *hGGPS* coding sequence, or both.

More particularly, the present invention relates to expression vectors which include nucleic acids encoding the hGGPS protein of the amino acid sequence of SEQ ID No 4 described therein under the control of either one regulatory sequence selected among the *hGGPS* regulatory polynucleotides, or alternatively under the control of an exogenous regulatory sequence.

A recombinant expression vector comprising a nucleic acid selected from the group consisting of the 5' or 3' regulatory regions of *hGGPPS*, or biologically active fragments or variants thereof, is also part of the present invention.

Generally, a recombinant vector of the invention may comprise any of the polynucleotides described herein, including regulatory sequences, and coding sequences, as well as any *hGGPPS* primer or probe as defined above. More particularly, the recombinant vectors of the present invention can comprise any of the polynucleotides described in the "*hGGPPS* cDNA Sequences" section, the "Coding Regions" section, "Genomic sequences" section and the "Oligonucleotide Probes And Primers" section.

Some of the elements which can be found in the vectors of the present invention are described in further detail in the following sections.

a) Vectors

A recombinant vector according to the invention comprises, but is not limited to, a YAC (Yeast Artificial Chromosome), a BAC (Bacterial Artificial Chromosome), a phage, a phagemid, a cosmid, a plasmid or even a linear DNA molecule which may consist of a chromosomal, non-chromosomal and synthetic DNA. Such a recombinant vector can comprise a transcriptional unit comprising an assembly of :

(1) a genetic element or elements having a regulatory role in gene expression, for example promoters or enhancers. Enhancers are cis-acting elements of DNA, usually from about 10 to 300 bp in length that act on the promoter to increase the transcription.

(2) a structural or coding sequence which is transcribed into mRNA and eventually translated into a polypeptide, and

(3) appropriate transcription initiation and termination sequences. Structural units intended for use in yeast or eukaryotic expression systems preferably include a leader sequence enabling extracellular secretion of translated protein by a host cell. Alternatively, where recombinant protein is expressed without a leader or transport sequence, it may include an N-terminal residue. This residue may or may not be subsequently cleaved from the expressed recombinant protein to provide a final product.

Generally, recombinant expression vectors will include origins of replication, selectable markers permitting transformation of the host cell, and a promoter derived from a highly expressed gene to direct transcription of a downstream structural sequence. The heterologous structural sequence is assembled in appropriate phase with translation initiation and termination sequences, and preferably a leader sequence capable of directing secretion of translated protein into the periplasmic space or extracellular medium.

The selectable marker genes for selection of transformed host cells are preferably dihydrofolate reductase or neomycin resistance for eukaryotic cell culture, TRP1 for *S. cerevisiae* or tetracycline, rifampicin or ampicillin resistance in *E. coli*, or levan saccharase for mycobacteria.

As a representative but non-limiting example, useful expression vectors for bacterial use can comprise a selectable marker and bacterial origin of replication derived from commercially available plasmids comprising genetic elements of pBR322 (ATCC 37017). Such commercial vectors include, for example, pKK223-3 (Pharmacia, Uppsala, Sweden), and GEM1 (Promega Biotec, Madison, WI, USA).

Large numbers of suitable vectors and promoters are known to those of skill in the art, and commercially available, such as bacterial vectors : pQE70, pQE60, pQE-9 (Qiagen), pbs, pD10, phagescript, psiX174, pbluescript SK, pbsks, pNH8A, pNH16A, pNH18A, pNH46A (Stratagene); ptrc99a, pKK223-3, pKK233-3, pDR540, pRIT5 (Pharmacia); or eukaryotic vectors : pWLNEO, pSV2CAT, pOG44, pXT1, pSG (Stratagene); pSVK3, pBPV, pMSG, pSVL (Pharmacia); baculovirus transfer vector pVL1392/1393 (Pharmingen); pQE-30 (QIAexpress).

A suitable vector for the expression of the hGGPS polypeptide of SEQ ID No 4 is a baculovirus vector that can be propagated in insect cells and in insect cell lines. A specific suitable host vector system is the pVL1392/1393 baculovirus transfer vector (Pharmingen) that is used to transfect the SF9 cell line (ATCC N°CRL 1711) which is derived from *Spodoptera frugiperda*.

Other suitable vectors for the expression of the hGGPS polypeptide of SEQ ID No 4 in a baculovirus expression system include those described by Chai et al. (1993), Vlasak et al. (1983) and Lenhard et al. (1996).

Mammalian expression vectors will comprise an origin of replication, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation signal, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking nontranscribed sequences. DNA sequences derived from the SV40 viral genome, for example SV40 origin, early promoter, enhancer, splice and polyadenylation signals may be used to provide the required nontranscribed genetic elements.

b) Promoters

The suitable promoter regions used in the expression vectors according to the present invention are chosen taking into account the cell host in which the heterologous gene has to be expressed.

A suitable promoter may be heterologous with respect to the nucleic acid for which it controls the expression or alternatively can be endogenous to the native polynucleotide containing the coding sequence to be expressed. Additionally, the promoter is generally heterologous with respect to the recombinant vector sequences within which the construct promoter/coding sequence has been inserted.

Preferred bacterial promoters are the LacI, LacZ, the T3 or T7 bacteriophage RNA polymerase promoters, the polyhedrin promoter, or the p10 protein promoter from baculovirus (Kit Novagen) (Smith et al., 1983; O'Reilly et al., 1992), the lambda P_R promoter or also the trc promoter.

Promoter regions can be selected from any desired gene using, for example, CAT (chloramphenicol transferase) vectors and more preferably pKK232-8 and pCM7 vectors. Particularly preferred bacterial promoters include lacI, lacZ, T3, T7, gpt, lambda PR, PL and trp. Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-L. Selection of a convenient vector and promoter is well within the level of ordinary skill in the art.

The choice of a promoter is well within the ability of a person skilled in the field of genetic egineering. For example, one may refer to the book of Sambrook et al. (1989) or also to the procedures described by Fuller et al. (1996).

The vector containing the appropriate DNA sequence as described above, more preferably a *hGGPS* gene regulatory polynucleotide, a polynucleotide encoding the hGGPS polypeptide of SEQ

ID No 4 or both of them, can be utilized to transform an appropriate host to allow the expression of the desired polypeptide or polynucleotide.

c) Other types of vectors

The *in vivo* expression of a hGGPS polypeptide of SEQ ID No 4 may be useful in order to correct a genetic defect related to the expression of the native gene in a host organism or to the production of a biologically inactive hGGPS protein.

Consequently, the present invention also deals with recombinant expression vectors mainly designed for the *in vivo* production of the hGGPS polypeptide of SEQ ID No 4 by the introduction of the appropriate genetic material in the organism of the patient to be treated. This genetic material may be introduced *in vitro* in a cell that has been previously extracted from the organism, the modified cell being subsequently reintroduced in the said organism, directly *in vivo* into the appropriate tissue, and preferably in the olfactory epithelium.

By « vector » according to this specific embodiment of the invention is intended either a circular or a linear DNA molecule.

One specific embodiment for a method for delivering a protein or peptide to the interior of a cell of a vertebrate *in vivo* comprises the step of introducing a preparation comprising a physiologically acceptable carrier and a naked polynucleotide operatively coding for the polypeptide of interest into the interstitial space of a tissue comprising the cell, whereby the naked polynucleotide is taken up into the interior of the cell and has a physiological effect.

In a specific embodiment, the invention provides a composition for the *in vivo* production of the hGGPS protein or polypeptide described herein. It comprises a naked polynucleotide operatively coding for this polypeptide, in solution in a physiologically acceptable carrier, and suitable for introduction into a tissue to cause cells of the tissue to express the said protein or polypeptide.

Compositions comprising a polynucleotide are described in the PCT application N° WO 90/11092 (Vical Inc.) and also in the PCT application N° WO 95/11307 (Institut Pasteur, INSERM, Université d'Ottawa) as well as in the articles of Tacson et al. (1996) and of Huygen et al. (1996).

The amount of the vector to be injected to the desired host organism vary according to the site of injection. As an indicative dose, it will be injected between 0,1 and 100 µg of the vector in an animal body, preferably a mammal body, for example a mouse body.

In another embodiment of the vector according to the invention, it may be introduced in vitro in a host cell, preferably in a host cell previously harvested from the animal to be treated and more preferably a somatic cell such as a muscle cell. In a subsequent step, the cell that has been transformed with the vector coding for the desired hGGPS polypeptide or the desired C-terminal fragment thereof is reintroduced into the animal body in order to deliver the recombinant protein within the body either locally or systemically.

In one specific embodiment, the vector is derived from an adenovirus. Preferred adenovirus vectors according to the invention are those described by Feldman and Steg (1996) or Ohno et al.

(1994). Another preferred recombinant adenovirus according to this specific embodiment of the present invention is the human adenovirus type 2 or 5 (Ad 2 or Ad 5) or an adenovirus of animal origin (French patent application N° FR-93.05954).

Retrovirus vectors and adeno-associated virus vectors are generally understood to be the recombinant gene delivery system of choice for the transfer of exogenous polynucleotides *in vivo* , particularly to mammals, including humans. These vectors provide efficient delivery of genes into cells, and the transferred nucleic acids are stably integrated into the chromosomal DNA of the host

Particularly preferred retroviruses for the preparation or construction of retroviral *in vitro* or *in vitro* gene delivery vehicles of the present invention include retroviruses selected from the group consisting of Mink-Cell Focus Inducing Virus, Murine Sarcoma Virus, Reticuloendotheliosis virus and Rous Sarcoma virus. Particularly preferred Murine Leukemia Viruses include the 4070A and the 1504A viruses, Abelson (ATCC No VR-999), Friend (ATCC No VR-245), Gross (ATCC No VR-590), Rauscher (ATCC No VR-998) and Moloney Murine Leukemia Virus (ATCC No VR-190; PCT Application No WO 94/24298). Particularly preferred Rous Sarcoma Viruses include Bryan high titer (ATCC Nos VR-334, VR-657, VR-726, VR-659 and VR-728). Other preferred retroviral vectors are those described in Roth et al. (1996), the PCT Application No WO 93/25234, the PCT Application No WO 94/ 06920, Roux et al., 1989, Julan et al., 1992 and Neda et al., 1991.

Yet another viral vector system that is contemplated by the invention consists in the adeno-associated virus (AAV). The adeno-associated virus is a naturally occurring defective virus that requires another virus, such as an adenovirus or a herpes virus, as a helper virus for efficient replication and a productive life cycle (Muzyczka et al., 1992). It is also one of the few viruses that may integrate its DNA into non-dividing cells, and exhibits a high frequency of stable integration (Flotte et al., 1992; Samulski et al., 1989; McLaughlin et al., 1989). One advantageous feature of AAV derives from its reduced efficacy for transducing primary cells relative to transformed cells.

Other compositions containing a vector of the invention advantageously comprise an oligonucleotide fragment of a nucleic sequence selected from the group consisting of SEQ ID Nos 2 or 3 as an antisense tool that inhibits the expression of the corresponding *hGGPS* gene. Preferred methods using antisense polynucleotide according to the present invention are the procedures described by Sczakiel et al. (1995) or also in the PCT Application No WO 95/24223.

Preferably, the antisense tools are chosen among the polynucleotides (15-200 bp long) that are complementary to the 5'end of the *hGGPS* mRNAs. In another embodiment, a combination of different antisense polynucleotides complementary to different parts of the desired targeted gene are used.

Preferred antisense polynucleotides according to the present invention are complementary to a sequence of the mRNAs of *hGGPS* that contains the translation initiation codon ATG.

Host cells

Another object of the invention consists in cell host that have been transformed or transfected with one of the polynucleotides described therein, and more precisely a polynucleotide either comprising a *hGGPS* regulatory polynucleotide or the coding sequence of the hGGPS polypeptide having the amino acid sequence of SEQ ID No 4. Are included cell hosts that are transformed (prokaryotic cells) or that are transfected (eukaryotic cells) with a recombinant vector such as those described above.

A cell host according to the present invention is characterized in that its genome or genetic background (including chromosome, plasmids) is modified by the heterologous nucleic acid coding for the hGGPS polypeptide of SEQ ID No 4.

More particularly, the cell hosts of the present invention can comprise any of the polynucleotides described in "*hGGPPS* cDNA Sequences" section, the "Coding Regions" section, "Genomic sequences" section and the "Oligonucleotide Probes And Primers" section.

Preferred cell hosts used as recipients for the expression vectors of the invention are the following :

a) Prokaryotic host cells : *Escherichia coli* strains (I.E. DH5-α strain) or *Bacillus subtilis*.

b) Eukaryotic host cells : HeLa cells (ATCC N°CCL2; N°CCL2.1; N°CCL2.2), Cv 1 cells (ATCC N°CCL70), COS cells (ATCC N°CRL1650; N°CRL1651), Sf-9 cells (ATCC N°CRL1711).

The constructs in the host cells can be used in a conventional manner to produce the gene product encoded by the recombinant sequence.

Following transformation of a suitable host and growth of the host to an appropriate cell density, the selected promoter is induced by appropriate means, such as temperature shift or chemical induction, and cells are cultivated for an additional period.

Cells are typically harvested by centrifugation, disrupted by physical or chemical means, and the resulting crude extract retained for further purification.

Microbial cells employed in expression of proteins can be disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents. Such methods are well known by the skill artisan.

Cell hosts can be used to generate transgenic animals. Therefore, the invention concerns a non-human host animal or mammal comprising a recombinant vector or a host cell according to the invention. More particularly, the invention concerns a mammalian host cell or a non-human host mammal comprising a *hGGPPS* gene disrupted by homologous recombination with a knock out vector and comprising a polynucleotide according to the invention.

hGGPPS Proteins and Polypeptide Fragments:

The term "hGGPPS polypeptides" is used herein to embrace all of the proteins and polypeptides of the present invention. Also forming part of the invention are polypeptides encoded

by the polynucleotides of the invention, as well as fusion polypeptides comprising such polypeptides. The invention embodies hGGPPS proteins from humans, including isolated or purified hGGPPS proteins consisting, consisting essentially, or comprising the sequence of SEQ ID No 4. It should be noted the hGGPPS proteins of the invention are based on the naturally-occurring variant of the amino acid sequence of human hGGPPS, wherein a phenylalanine residue is at positions 204, 257, 295 of SEQ ID No 4, a cysteine residue is at position 205 of SEQ ID No 4, a proline residue is at position 225 of SEQ ID No 4, and a glutamic acid residue is at position 252 of SEQ ID No 4.

The present invention embodies isolated, purified, and recombinant polypeptides comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 4, wherein said contiguous span includes at least one amino acid selected from the group consisting of a Phe at positions 204, 257, 295 of SEQ ID No 4, a Cys at position 205 of SEQ ID No 4, a Pro at position 225 of SEQ ID No 4, and a Glu at position 252 of SEQ ID No 4. In other preferred embodiments the contiguous stretch of amino acids comprises the site of a mutation or functional mutation, including a deletion, addition, swap or truncation of the amino acids in the hGGPPS protein sequence.

hGGPPS proteins are preferably isolated from human or mammalian tissue samples or expressed from human or mammalian genes. The hGGPPS polypeptides of the invention can be made using routine expression methods known in the art. The polynucleotide encoding the desired polypeptide, is ligated into an expression vector suitable for any convenient host. Both eukaryotic and prokaryotic host systems is used in forming recombinant polypeptides, and a summary of some of the more common systems. The polypeptide is then isolated from lysed cells or from the culture medium and purified to the extent needed for its intended use. Purification is by any technique known in the art, for example, differential extraction, salt fractionation, chromatography, centrifugation, and the like. See, for example, Methods in Enzymology for a variety of methods for purifying proteins.

In addition, shorter protein fragments is produced by chemical synthesis. Alternatively the proteins of the invention is extracted from cells or tissues of humans or non-human animals. Methods for purifying proteins are known in the art, and include the use of detergents or chaotropic agents to disrupt particles followed by differential extraction and separation of the polypeptides by ion exchange chromatography, affinity chromatography, sedimentation according to density, and gel electrophoresis.

Any *hGGPPS* cDNA, including SEQ ID Nos 2 and 3, is used to express hGGPPS proteins and polypeptides. The nucleic acid encoding the hGGPPS protein or polypeptide to be expressed is operably linked to a promoter in an expression vector using conventional cloning technology. The hGGPPS insert in the expression vector may comprise the full coding sequence for the hGGPPS protein or a portion thereof. For example, the hGGPPS derived insert may encode a polypeptide comprising at

least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 consecutive amino acids of the hGGPPS protein of SEQ ID No 4, wherein said consecutive amino acids comprise at least one amino acid selected from the group consisting of a Phe at positions 204, 257, 295 of SEQ ID No 4, a Cys at position 205 of SEQ ID No 4, a Pro at position 225 of SEQ ID No 4, and a Glu at position 252 of SEQ ID No 4.

The expression vector is any of the mammalian, yeast, insect or bacterial expression systems known in the art. Commercially available vectors and expression systems are available from a variety of suppliers including Genetics Institute (Cambridge, MA), Stratagene (La Jolla, California), Promega (Madison, Wisconsin), and Invitrogen (San Diego, California). If desired, to enhance expression and facilitate proper protein folding, the codon context and codon pairing of the sequence is optimized for the particular expression organism in which the expression vector is introduced, as explained by Hatfield, et al., U.S. Patent No. 5,082,767, the disclosures of which are incorporated by reference herein in their entirety.

In one embodiment, the entire coding sequence of the hGGPPS cDNA through the poly A signal of the cDNA are operably linked to a promoter in the expression vector. Alternatively, if the nucleic acid encoding a portion of the hGGPPS protein lacks a methionine to serve as the initiation site, an initiating methionine can be introduced next to the first codon of the nucleic acid using conventional techniques. Similarly, if the insert from the hGGPPS cDNA lacks a poly A signal, this sequence can be added to the construct by, for example, splicing out the Poly A signal from pSG5 (Stratagene) using BglI and SalI restriction endonuclease enzymes and incorporating it into the mammalian expression vector pXT1 (Stratagene). pXT1 contains the LTRs and a portion of the gag gene from Moloney Murine Leukemia Virus. The position of the LTRs in the construct allow efficient stable transfection. The vector includes the Herpes Simplex Thymidine Kinase promoter and the selectable neomycin gene. The nucleic acid encoding the hGGPPS protein or a portion thereof is obtained by PCR from a bacterial vector containing the hGGPPS cDNA of SEQ ID Nos 2 and 3 using oligonucleotide primers complementary to the hGGPPS cDNA or portion thereof and containing restriction endonuclease sequences for Pst I incorporated into the 5'primer and BglII at the 5' end of the corresponding cDNA 3' primer, taking care to ensure that the sequence encoding the hGGPPS protein or a portion thereof is positioned properly with respect to the poly A signal. The purified fragment obtained from the resulting PCR reaction is digested with PstI, blunt ended with an exonuclease, digested with Bgl II, purified and ligated to pXT1, now containing a poly A signal and digested with BglII.

The ligated product is transfected into mouse NIH 3T3 cells using Lipofectin (Life Technologies, Inc., Grand Island, New York) under conditions outlined in the product specification. Positive transfectants are selected after growing the transfected cells in 600ug/ml G418 (Sigma, St. Louis, Missouri).

The above procedures may also be used to express a mutant hGGPPS protein responsible for a detectable phenotype or a portion thereof.

The expressed protein is purified using conventional purification techniques such as ammonium sulfate precipitation or chromatographic separation based on size or charge. The protein encoded by the nucleic acid insert may also be purified using standard immunochromatography techniques. In such procedures, a solution containing the expressed hGGPPS protein or portion thereof, such as a cell extract, is applied to a column having antibodies against the hGGPPS protein or portion thereof is attached to the chromatography matrix. The expressed protein is allowed to bind the immunochromatography column. Thereafter, the column is washed to remove non-specifically bound proteins. The specifically bound expressed protein is then released from the column and recovered using standard techniques.

To confirm expression of the hGGPPS protein or a portion thereof, the proteins expressed from host cells containing an expression vector containing an insert encoding the hGGPPS protein or a portion thereof can be compared to the proteins expressed in host cells containing the expression vector without an insert. The presence of a band in samples from cells containing the expression vector with an insert which is absent in samples from cells containing the expression vector without an insert indicates that the hGGPPS protein or a portion thereof is being expressed. Generally, the band will have the mobility expected for the hGGPPS protein or portion thereof. However, the band may have a mobility different than that expected as a result of modifications such as glycosylation, ubiquitination, or enzymatic cleavage.

Antibodies capable of specifically recognizing the expressed hGGPPS protein or a portion thereof are described below.

If antibody production is not possible, the nucleic acids encoding the hGGPPS protein or a portion thereof is incorporated into expression vectors designed for use in purification schemes employing chimeric polypeptides. In such strategies the nucleic acid encoding the hGGPPS protein or a portion thereof is inserted in frame with the gene encoding the other half of the chimera. The other half of the chimera is β-globin or a nickel binding polypeptide encoding sequence. A chromatography matrix having antibody to β-globin or nickel attached thereto is then used to purify the chimeric protein. Protease cleavage sites is engineered between the β-globin gene or the nickel binding polypeptide and the hGGPPS protein or portion thereof. Thus, the two polypeptides of the chimera is separated from one another by protease digestion.

One useful expression vector for generating β-globin chimeric proteins is pSG5 (Stratagene), which encodes rabbit β-globin. Intron II of the rabbit β-globin gene facilitates splicing of the expressed transcript, and the polyadenylation signal incorporated into the construct increases the level of expression. These techniques are well known to those skilled in the art of molecular biology. Standard methods are published in methods texts such as Davis et al., (1986) and many of the methods are available from Stratagene, Life Technologies, Inc., or Promega. Polypeptide may additionally be produced from the construct using in vitro translation systems such as the In vitro Express™ Translation Kit (Stratagene).

Antibodies That Bind hGGPPS Polypeptides of the Invention

Any hGGPPS polypeptide or whole protein may be used to generate antibodies capable of specifically binding to an expressed hGGPPS protein or fragments thereof as described.

One antibody composition of the invention is capable of specifically binding or specifically bind to the variant of the hGGPPS protein of SEQ ID No 4. For an antibody composition to specifically bind to a first variant of hGGPPS, it must demonstrate at least a 5%, 10%, 15%, 20%, 25%, 50%, or 100% greater binding affinity for a full length first variant of the hGGPPS protein than for a full length second variant of the hGGPPS protein in an ELISA, RIA, or other antibody-based binding assay.

In a preferred embodiment of polyclonal or monoclonal antibodies of the invention consists in antibodies raised against a C-terminal portion of the hGGPS polypeptide of the amino acid sequence of SEQ ID No 4, more preferably antibodies raise against a peptide fragment of the hGGPS polypeptide having the amino acid sequence starting from the amino acid at position 200 and ending at the amino acid in position 300 of the hGGPS polypeptide of SEQ ID No 4, or peptide fragments thereof.

In a preferred embodiment, the invention concerns antibody compositions, either polyclonal or monoclonal, capable of selectively binding, or selectively bind to an epitope-containing a polypeptide comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 4 , wherein said epitope comprises at least one amino acid selected from the group consisting of a Phe at positions 204, 257, 295 of SEQ ID No 4, a Cys at position 205 of SEQ ID No 4, a Pro at position 225 of SEQ ID No 4, and a Glu at position 252 of SEQ ID No 4.

The invention also concerns a purified or isolated antibody capable of specifically binding to a mutated hGGPPS protein or to a fragment or variant thereof comprising an epitope of the mutated hGGPPS protein.

In a preferred embodiment, the invention concerns the use in the manufacture of antibodies of a polypeptide comprising a contiguous span of at least 6 amino acids, preferably at least 8 to 10 amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 amino acids of SEQ ID No 4 , wherein said epitope comprises at least one amino acid selected from the group consisting of a Phe at positions 204, 257, 295 of SEQ ID No 4, a Cys at position 205 of SEQ ID No 4, a Pro at position 225 of SEQ ID No 4, and a Glu at position 252 of SEQ ID No 4.

Non-human animals or mammals, whether wild-type or transgenic, which express a different species of hGGPPS than the one to which antibody binding is desired, and animals which do not express hGGPPS (i.e. a hGGPPS knock out animal as described herein) are particularly useful for preparing antibodies. hGGPPS knock out animals will recognize all or most of the exposed regions of a hGGPPS protein as foreign antigens, and therefore produce antibodies with a wider array of hGGPPS epitopes. Moreover, smaller polypeptides with only 10 to 30 amino acids may be useful in

obtaining specific binding to any one of the hGGPPS proteins. In addition, the humoral immune system of animals which produce a species of hGGPPS that resembles the antigenic sequence will preferentially recognize the differences between the animal's native hGGPPS species and the antigen sequence, and produce antibodies to these unique sites in the antigen sequence. Such a technique will be particularly useful in obtaining antibodies that specifically bind to any one of the hGGPPS proteins.

Antibody preparations prepared according to either protocol are useful in quantitative immunoassays which determine concentrations of antigen-bearing substances in biological samples; they are also used semi-quantitatively or qualitatively to identify the presence of antigen in a biological sample. The antibodies may also be used in therapeutic compositions for killing cells expressing the protein or reducing the levels of the protein in the body.

The antibodies of the invention may be labeled by any one of the radioactive, fluorescent or enzymatic labels known in the art.

Consequently, the invention is also directed to a method for detecting specifically the presence of a hGGPPS polypeptide according to the invention in a biological sample, said method comprising the following steps :

a) bringing into contact the biological sample with a polyclonal or monoclonal antibody that specifically binds a hGGPPS polypeptide comprising an amino acid sequence of SEQ ID No 4, or to a peptide fragment or variant thereof; and

b) detecting the antigen-antibody complex formed.

The invention also concerns a diagnostic kit for detecting *in vitro* the presence of a hGGPPS polypeptide according to the present invention in a biological sample, wherein said kit comprises:

a) a polyclonal or monoclonal antibody that specifically binds a hGGPPS polypeptide comprising an amino acid sequence of SEQ ID No 4, or to a peptide fragment or variant thereof, optionally labeled;

b) a reagent allowing the detection of the antigen-antibody complexes formed, said reagent carrying optionally a label, or being able to be recognized itself by a labeled reagent, more particularly in the case when the above-mentioned monoclonal or polyclonal antibody is not labeled by itself.

Method For Screening Ligands That Modulate The Expression Of The *hGGPPS* Gene.

Another subject of the present invention is a method for screening molecules that modulate the expression of the hGGPPS protein. Such a screening method comprises the steps of:

a) cultivating a prokaryotic or an eukaryotic cell that has been transfected with a nucleotide sequence encoding the hGGPPS protein or a variant or a fragment thereof, placed under the control of its own promoter;

b) bringing into contact the cultivated cell with a molecule to be tested;

c) quantifying the expression of the hGGPPS protein or a variant or a fragment thereof.

In an embodiment, the nucleotide sequence encoding the hGGPPS protein or a variant or a fragment thereof, preferably a fragment comprising an allele of the biallelic marker 5-187-77, and the complement thereof.

In one embodiment of the invention, the method for the screening of a candidate substance or molecule modulating the expression of the *hGGPS* genecomprises the following steps :

a) providing a recombinant host cell expressing a nucleic acid, wherein said nucleic acid comprises a nucleotide sequence selected from the group consisting of SEQ ID Nos 1, 2 and 3 or a fragment thereof;

b) obtaining a candidate substance, and

c) determining the ability of the candidate substance to modulate the expression levels of the nucleotide sequence selected from the group consisting of SEQ ID Nos 1, 2 and 3 or a fragment thereof.

Using DNA recombination techniques well known by the one skill in the art, the hGGPPS protein encoding DNA sequence is inserted into an expression vector, downstream from its promoter sequence. As an illustrative example, the promoter sequence of the *hGGPPS* gene is contained in the nucleic acid of the 5' regulatory region.

The quantification of the expression of the hGGPPS protein may be realized either at the mRNA level or at the protein level. In the latter case, polyclonal or monoclonal antibodies may be used to quantify the amounts of the hGGPPS protein that have been produced, for example in an ELISA or a RIA assay.

In a preferred embodiment, the quantification of the *hGGPPS* mRNA is realized by a quantitative PCR amplification of the cDNA obtained by a reverse transcription of the total mRNA of the cultivated *hGGPPS* -transfected host cell, using a pair of primers specific for *hGGPPS*.

The present invention also concerns a method for screening substances or molecules that are able to increase, or in contrast to decrease, the level of expression of the *hGGPPS* gene. Such a method may allow the one skilled in the art to select substances exerting a regulating effect on the expression level of the *hGGPPS* gene and which may be useful as active ingredients included in pharmaceutical compositions.

Thus, is also part of the present invention a method for screening of a candidate substance or molecule that modulated the expression of the *hGGPPS* gene, this method comprises the following steps:

- providing a recombinant cell host containing a nucleic acid, wherein said nucleic acid comprises a nucleotide sequence of the 5' regulatory region or a biologically active fragment or variant thereof located upstream a polynucleotide encoding a detectable protein;

- obtaining a candidate substance; and

- determining the ability of the candidate substance to modulate the expression levels of the polynucleotide encoding the detectable protein.

In a further embodiment, the nucleic acid comprising the nucleotide sequence of the 5' regulatory region or a biologically active fragment or variant thereof also includes a 5'UTR region of the *hGGPPS* cDNA of SEQ ID Nos 2 or 3, or one of its biologically active fragments or variants thereof.

Among the preferred polynucleotides encoding a detectable protein, there may be cited polynucleotides encoding beta galactosidase, green fluorescent protein (GFP) and chloramphenicol acetyl transferase (CAT).

The invention also pertains to kits useful for performing the herein described screening method. Preferably, such kits comprise a recombinant vector that allows the expression of a nucleotide sequence of the 5' regulatory region or a biologically active fragment or variant thereof located upstream and operably linked to a polynucleotide encoding a detectable protein or the hGGPPS protein or a fragment or a variant thereof.

In another embodiment of a method for the screening of a candidate substance or molecule that modulates the expression of the *hGGPPS* gene, wherein said method comprises the following steps:

a) providing a recombinant host cell containing a nucleic acid, wherein said nucleic acid comprises a 5'UTR sequence of the *hGGPPS* cDNA of SEQ ID Nos 2 or 3, or one of its biologically active fragments or variants, the 5'UTR sequence or its biologically active fragment or variant being operably linked to a polynucleotide encoding a detectable protein;

b) obtaining a candidate substance; and

c) determining the ability of the candidate substance to modulate the expression levels of the polynucleotide encoding the detectable protein.

In a specific embodiment of the above screening method, the nucleic acid that comprises a nucleotide sequence selected from the group consisting of the 5'UTR sequence of the *hGGPPS* cDNA of SEQ ID Nos 2 or 3 or one of its biologically active fragments or variants, includes a promoter sequence which is endogenous with respect to the *hGGPPS* 5'UTR sequence.

In another specific embodiment of the above screening method, the nucleic acid that comprises a nucleotide sequence selected from the group consisting of the 5'UTR sequence of the *hGGPPS* cDNA of SEQ ID Nos 2 or 3 or one of its biologically active fragments or variants, includes a promoter sequence which is exogenous with respect to the *hGGPPS* 5'UTR sequence defined therein.

In a further preferred embodiment, the nucleic acid comprising the 5'-UTR sequence of the *hGGPPS* cDNA or SEQ ID Nos 2 or 3 or the biologically active fragments thereof, preferably those including the biallelic marker 5-187-77 or the complement thereof.

The invention further deals with a kit for the screening of a candidate substance modulating the expression of the *hGGPPS* gene, wherein said kit comprises a recombinant vector that comprises a nucleic acid including a 5'UTR sequence of the *hGGPPS* cDNA of SEQ ID Nos 2 or 3, or one of their biologically active fragments or variants, the 5'UTR sequence or its biologically active fragment or variant being operably linked to a polynucleotide encoding a detectable protein.

For the design of suitable recombinant vectors useful for performing the screening methods described above, it will be referred to the section of the present specification wherein the preferred recombinant vectors of the invention are detailed.

Expression levels and patterns of *hGGPPS* may be analyzed by solution hybridization with long probes as described in International Patent Application No. WO 97/05277, the entire contents of which are incorporated herein by reference. Briefly, the *hGGPPS* cDNA or the *hGGPPS* genomic DNA described above, or fragments thereof, is inserted at a cloning site immediately downstream of a bacteriophage (T3, T7 or SP6) RNA polymerase promoter to produce antisense RNA. Preferably, the *hGGPPS* insert comprises at least 100 or more consecutive nucleotides of the genomic DNA sequence or the cDNA sequences. The plasmid is linearized and transcribed in the presence of ribonucleotides comprising modified ribonucleotides (i.e. biotin-UTP and DIG-UTP). An excess of this doubly labeled RNA is hybridized in solution with mRNA isolated from cells or tissues of interest. The hybridization is performed under standard stringent conditions (40-50°C for 16 hours in an 80% formamide, 0. 4 M NaCl buffer, pH 7-8). The unhybridized probe is removed by digestion with ribonucleases specific for single-stranded RNA (i.e. RNases CL3, T1, Phy M, U2 or A). The presence of the biotin-UTP modification enables capture of the hybrid on a microtitration plate coated with streptavidin. The presence of the DIG modification enables the hybrid to be detected and quantified by ELISA using an anti-DIG antibody coupled to alkaline phosphatase.

Quantitative analysis of *hGGPPS* gene expression may also be performed using arrays. As used herein, the term array means a one dimensional, two dimensional, or multidimensional arrangement of a plurality of nucleic acids of sufficient length to permit specific detection of expression of mRNAs capable of hybridizing thereto. For example, the arrays may contain a plurality of nucleic acids derived from genes whose expression levels are to be assessed. The arrays may include the *hGGPPS* genomic DNA, the *hGGPPS* cDNA sequences or the sequences complementary thereto or fragments thereof, particularly those comprising the biallelic marker 5-187-77. Preferably, the fragments are at least 15 nucleotides in length. In other embodiments, the fragments are at least 25 nucleotides in length. In some embodiments, the fragments are at least 50 nucleotides in length. More preferably, the fragments are at least 100 nucleotides in length. In another preferred embodiment, the fragments are more than 100 nucleotides in length. In some embodiments the fragments may b more than 500 nucleotides in length.

For example, quantitative analysis of *hGGPPS* gene expression may be performed with a complementary DNA microarray as described by Schena et al.(1995 and 1996). Full length *hGGPPS* cDNAs or fragments thereof are amplified by PCR and arrayed from a 96-well microtiter plate onto silylated microscope slides using high-speed robotics. Printed arrays are incubated in a humid chamber to allow rehydration of the array elements and rinsed, once in 0. 2% SDS for 1 min, twice in water for 1 min and once for 5 min in sodium borohydride solution. The arrays are submerged in water for 2 min at 95°C, transferred into 0. 2% SDS for 1 min, rinsed twice with water, air dried and stored in the dark at 25°C.

Cell or tissue mRNA is isolated or commercially obtained and probes are prepared by a single round of reverse transcription. Probes are hybridized to 1 cm^2 microarrays under a 14 x 14 mm glass coverslip for 6-12 hours at 60°C. Arrays are washed for 5 min at 25°C in low stringency wash buffer (1 x SSC/0. 2% SDS), then for 10 min at room temperature in high stringency wash buffer (0. 1 x SSC/0. 2% SDS). Arrays are scanned in 0. 1 x SSC using a fluorescence laser scanning device fitted with a custom filter set. Accurate differential expression measurements are obtained by taking the average of the ratios of two independent hybridizations.

Quantitative analysis of *hGGPPS* gene expression may also be performed with full length *hGGPPS* cDNAs or fragments thereof in complementary DNA arrays as described by Pietu et al.(1996). The full length *hGGPPS* cDNA or fragments thereof is PCR amplified and spotted on membranes. Then, mRNAs originating from various tissues or cells are labeled with radioactive nucleotides. After hybridization and washing in controlled conditions, the hybridized mRNAs are detected by phospho-imaging or autoradiography. Duplicate experiments are performed and a quantitative analysis of differentially expressed mRNAs is then performed.

Alternatively, expression analysis using the *hGGPPS* genomic DNA, the *hGGPPS* cDNA, or fragments thereof can be done through high density nucleotide arrays as described by Lockhart et al.(1996) and Sosnowsky et al.(1997). Oligonucleotides of 15-50 nucleotides from the sequences of the *hGGPPS* genomic DNA, the *hGGPPS* cDNA sequences particularly those comprising the biallelic marker 5-187-77, or the sequences complementary thereto, are synthesized directly on the chip (Lockhart et al., supra) or synthesized and then addressed to the chip (Sosnowski et al., supra). Preferably, the oligonucleotides are about 20 nucleotides in length.

hGGPPS cDNA probes labeled with an appropriate compound, such as biotin, digoxigenin or fluorescent dye, are synthesized from the appropriate mRNA population and then randomly fragmented to an average size of 50 to 100 nucleotides. The said probes are then hybridized to the chip. After washing as described in Lockhart et al., supra and application of different electric fields (Sosnowsky et al., 1997)., the dyes or labeling compounds are detected and quantified. Duplicate hybridizations are performed. Comparative analysis of the intensity of the signal originating from cDNA probes on the same target oligonucleotide in different cDNA samples indicates a differential expression of *hGGPPS* mRNA.

Throughout this application, various publications, patents and published patent applications are cited. The disclosures of these publications, patents and published patent specification referenced in this application are hereby incorporated by reference into the present disclosure to more fully describe the sate of the art to which this invention pertains.

EXAMPLES

Example 1 :

Analysis of the mRNAs encoding the hGGPS polypeptide of SEQ ID No 4 synthesized by the cells.

Human *GGPS* cDNA was obtained as follows : 4µl of ethanol suspension containing 1 mg of human prostate total RNA (Clontech laboratories, Inc., Palo Alto, USA; Catalogue N. 64038-1) was centrifuged, and the resulting pellet was air dried for 30 minutes at room temperature.

First strand cDNA synthesis was performed using the AdvantageTM RT-for- PCR kit (Clontech laboratories Inc., catalogue N. K1402-1). 1 µl of 20 mM solution of a specific oligo dT primer was added to 12.5 µl of RNA solution in water, heated at 74°C for 2.5 min and rapidly quenched in an ice bath. 10 µl of 5 x RT buffer (50 mM Tris-HCl, pH 8.3, 75 mM KCl, 3 mM $MgCl_2$), 2.5 µl of dNTP mix (10 mM each), 1.25 µl of human recombinant placental RNA inhibitor were mixed with 1 ml of MMLV reverse transcriptase (200 units). 6.5 µl of this solution were added to RNA-primer mix and incubated at 42°C for one hour. 80 µl of water were added and the solution was incubated at 94°C for 5 minutes.

5µl of the resulting solution were used in a Long Range PCR reaction with hot start, in 50 µl final volume, using 2 units of rtTHXL, 20 pmol/µl of each of 5'-TGGAGAAGACTCAAGAAACAGTCCAAA-3' (from the nucleotide in position 86 to the nucleotide in position 112 of SEQ ID No 1) and 5'-CCTGGAAGCAAGTCTTTTTTATTGACG-3' (from the nucleotide in position 1285 to the nucleotide in position 1311 of SEQ ID No 1) primers with 35 cycles of elongation for 6 minutes at 67°C in thermocycler.

The amplification products corresponding to both cDNA strands are partially sequenced in order to ensure the specificity of the amplification reaction.

Results of Northern blot analysis of prostate mRNAs support the existence of a *hGGPS* cDNA which corresponds to the nucleotide sequence of SEQ ID No 1.

Example 2 :

Detection of *hGGPS* biallelic markers: DNA extraction

Donors were unrelated and healthy. They presented a sufficient diversity for being representative of a French heterogeneous population. The DNA from 100 individuals was extracted and tested for the detection of the biallelic markers.

30 ml of peripheral venous blood were taken from each donor in the presence of EDTA. Cells (pellet) were collected after centrifugation for 10 minutes at 2000 rpm. Red cells were lysed by a lysis solution (50 ml final volume : 10 mM Tris pH7.6; 5 mM $MgCl_2$; 10 mM NaCl). The solution was centrifuged (10 minutes, 2000 rpm) as many times as necessary to eliminate the residual red cells present in the supernatant, after resuspension of the pellet in the lysis solution.

The pellet of white cells was lysed overnight at 42°C with 3.7 ml of lysis solution composed of:

- 3 ml TE 10-2 (Tris-HCl 10 mM, EDTA 2 mM) / NaCl 0.4 M
- 200 µl SDS 10%
- 500 µl K-proteinase (2 mg K-proteinase in TE 10-2 / NaCl 0.4 M).

For the extraction of proteins, 1 ml saturated NaCl (6M) (1/3.5 v/v) was added. After vigorous agitation, the solution was centrifuged for 20 minutes at 10000 rpm.

For the precipitation of DNA, 2 to 3 volumes of 100% ethanol were added to the previous supernatant, and the solution was centrifuged for 30 minutes at 2000 rpm. The DNA solution was rinsed three times with 70% ethanol to eliminate salts, and centrifuged for 20 minutes at 2000 rpm. The pellet was dried at 37°C, and resuspended in 1 ml TE 10-1 or 1 ml water. The DNA concentration was evaluated by measuring the OD at 260 nm (1 unit OD = 50 µg/ml DNA).

To determine the presence of proteins in the DNA solution, the OD 260 / OD 280 ratio was determined. Only DNA preparations having a OD 260 / OD 280 ratio between 1.8 and 2 were used in the subsequent examples described below.

The pool was constituted by mixing equivalent quantities of DNA from each individual.

Example 3 :

Detection of the biallelic markers: amplification of genomic DNA by PCR

The amplification of specific genomic sequences of the DNA samples of example 2 was carried out on the pool of DNA obtained previously. In addition, 50 individual samples were similarly amplified.

PCR assays were performed using the following protocol:

Final volume	25 µl
DNA	2 ng/µl
$MgCl_2$	2 mM

dNTP (each)	200 µM
primer (each)	2.9 ng/µl
Ampli Taq Gold DNA polymerase	0.05 unit/µl
PCR buffer (10x = 0.1 M TrisHCl pH8.3 0.5M KCl)	1x

Each pair of first primers was designed using the sequence information of the *hGGPS* gene disclosed herein and the OSP software (Hillier & Green, 1991). This first pair of primers was about 20 nucleotides in length and had the sequences disclosed in Table 1 in the columns labeled PU and RP.

Table 1

Amplicon	Position range of the amplicon in SEQ ID genomic	Position range of amplification primer in SEQ ID No genomic	Complementary position range of amplification primer in SEQ ID No genomic
5-187	13982-14409	13982-14000	14390-14409

The sequences of the amplification primers B1 and C1 are respectively disclosed in SEQ ID Nos 8 and 9.

Preferably, the primers contained a common oligonucleotide tail upstream of the specific bases targeted for amplification which was useful for sequencing. Primers PU contain the following additional PU 5' sequence : TGTAAAACGACGGCCAGT (SEQ ID No 10); primers RP contain the following RP 5' sequence : CAGGAAACAGCTATGACC (SEQ ID No 11).

The synthesis of these primers was performed following the phosphoramidite method, on a GENSET UFPS 24.1 synthesizer.

DNA amplification was performed on a Genius II thermocycler. After heating at 95°C for 10 min, 40 cycles were performed. Each cycle comprised: 30 sec at 95°C, 54°C for 1 min, and 30 sec at 72°C. For final elongation, 10 min at 72°C ended the amplification. The quantities of the amplification products obtained were determined on 96-well microtiter plates, using a fluorometer and Picogreen as intercalant agent (Molecular Probes).

Example 4 :

Detection of the biallelic markers: sequencing of amplified genomic DNA and identification of polymorphisms.

The sequencing of the amplified DNA obtained in example 3 was carried out on ABI 377 sequencers. The sequences of the amplification products were determined using automated dideoxy terminator sequencing reactions with a dye terminator cycle sequencing protocol. The products of the sequencing reactions were run on sequencing gels and the sequences were determined using gel image analysis.

The sequence data were further evaluated to detect the presence of biallelic markers among the pooled amplified fragments. The polymorphism search was based on the presence of

superimposed peaks in the electrophoresis pattern resulting from different bases occurring at the same position as described previously.

Table 2 shows the biallelic marker that has been detected after the sequence analysis of the amplification fragments generated by PCR.

Table 2

Amplicon	Marker Name	Localization in *hGGPPS* gene	Polymorphism	BM position in SEQ ID 1	Position of a probe in SEQ ID No 1
5-187	5-187-77	Intron 3	Insertion T	14058	14036-14081

The two alleles of the biallelic marker 5-187-77 can be defined by an oligonucleotide comprising the polymorphic base. The sequence of such oligonucleotides are disclosed in SEQ ID Nos 5 and 6.

Example 5 :

Validation of the polymorphisms through microsequencing

The biallelic marker identified in example 4 was further confirmed through microsequencing. Microsequencing was carried out for each individual DNA sample described in Example 2.

Amplification from genomic DNA of individuals was performed by PCR as described above for the detection of the biallelic markers with the same set of PCR primers (Table 1).

The preferred primers used in microsequencing were about 20 nucleotides in length and hybridized just upstream of the considered polymorphic base. According to the invention, the primer used in microsequencing is detailed in Table 3.

Table 3

Marker Name	Microsequencing primer
5-187-77	SEQ ID No 7

The microsequencing reaction was performed as follows :

After purification of the amplification products, the microsequencing reaction mixture was prepared by adding, in a 20µl final volume: 10 pmol microsequencing oligonucleotide, 1 U Thermosequenase (Amersham E79000G), 1.25 µl Thermosequenase buffer (260 mM Tris HCl pH 9.5, 65 mM $MgCl_2$), and the two appropriate fluorescent ddNTPs (Perkin Elmer, Dye Terminator Set 401095) complementary to the nucleotides at the polymorphic site of each biallelic marker tested, following the manufacturer's recommendations. After 4 minutes at 94°C, 20 PCR cycles of 15 sec at 55°C, 5 sec at 72°C, and 10 sec at 94°C were carried out in a Tetrad PTC-225 thermocycler (MJ Research). The unincorporated dye terminators were then removed by ethanol precipitation. Samples were finally resuspended in formamide-EDTA loading buffer and heated for 2 min at 95°C before

being loaded on a polyacrylamide sequencing gel. The data were collected by an ABI PRISM 377 DNA sequencer and processed using the GENESCAN software (Perkin Elmer).

Following gel analysis, data were automatically processed with software that allows the determination of the alleles of biallelic markers present in each amplified fragment.

The software evaluates such factors as whether the intensities of the signals resulting from the above microsequencing procedures are weak, normal, or saturated, or whether the signals are ambiguous. In addition, the software identifies significant peaks (according to shape and height criteria). Among the significant peaks, peaks corresponding to the targeted site are identified based on their position. When two significant peaks are detected for the same position, each sample is categorized classification as homozygous or heterozygous type based on the height ratio.

Example 6 :

Preparation of Antibody Compositions to the GENE protein

Substantially pure protein or polypeptide is isolated from transfected or transformed cells containing an expression vector encoding the hGGPPS protein or a portion thereof. The concentration of protein in the final preparation is adjusted, for example, by concentration on an Amicon filter device, to the level of a few micrograms/ml. Monoclonal or polyclonal antibody to the protein can then be prepared as follows:

A. Monoclonal Antibody Production by Hybridoma Fusion

Monoclonal antibody to epitopes in the hGGPPS protein or a portion thereof can be prepared from murine hybridomas according to the classical method of Kohler, G. and Milstein, C., (1975) or derivative methods thereof. Also see Harlow, E., and D. Lane. 1988..

Briefly, a mouse is repetitively inoculated with a few micrograms of the hGGPPS protein or a portion thereof over a period of a few weeks. The mouse is then sacrificed, and the antibody producing cells of the spleen isolated. The spleen cells are fused by means of polyethylene glycol with mouse myeloma cells, and the excess unfused cells destroyed by growth of the system on selective media comprising aminopterin (HAT media). The successfully fused cells are diluted and aliquots of the dilution placed in wells of a microtiter plate where growth of the culture is continued. Antibody-producing clones are identified by detection of antibody in the supernatant fluid of the wells by immunoassay procedures, such as ELISA, as originally described by Engvall, (1980), and derivative methods thereof. Selected positive clones can be expanded and their monoclonal antibody product harvested for use. Detailed procedures for monoclonal antibody production are described in Davis, L. et al. Basic Methods in Molecular Biology Elsevier, New York. Section 21-2.

B. Polyclonal Antibody Production by Immunization

Polyclonal antiserum containing antibodies to heterogeneous epitopes in the hGGPPS protein or a portion thereof can be prepared by immunizing suitable non-human animal with the hGGPPS protein or a portion thereof, which can be unmodified or modified to enhance

immunogenicity. A suitable non-human animal is preferably a non-human mammal is selected, usually a mouse, rat, rabbit, goat, or horse. Alternatively, a crude preparation which has been enriched for hGGPPS concentration can be used to generate antibodies. Such proteins, fragments or preparations are introduced into the non-human mammal in the presence of an appropriate adjuvant (e.g. aluminum hydroxide, RIBI, etc.) which is known in the art. In addition the protein, fragment or preparation can be pretreated with an agent which will increase antigenicity, such agents are known in the art and include, for example, methylated bovine serum albumin (mBSA), bovine serum albumin (BSA), Hepatitis B surface antigen, and keyhole limpet hemocyanin (KLH). Serum from the immunized animal is collected, treated and tested according to known procedures. If the serum contains polyclonal antibodies to undesired epitopes, the polyclonal antibodies can be purified by immunoaffinity chromatography.

Effective polyclonal antibody production is affected by many factors related both to the antigen and the host species. Also, host animals vary in response to site of inoculations and dose, with both inadequate or excessive doses of antigen resulting in low titer antisera. Small doses (ng level) of antigen administered at multiple intradermal sites appears to be most reliable. Techniques for producing and processing polyclonal antisera are known in the art, see for example, Mayer and Walker (1987). An effective immunization protocol for rabbits can be found in Vaitukaitis, J. et al. (1971).

Booster injections can be given at regular intervals, and antiserum harvested when antibody titer thereof, as determined semi-quantitatively, for example, by double immunodiffusion in agar against known concentrations of the antigen, begins to fall. See, for example, Ouchterlony, O. et al., (1973). Plateau concentration of antibody is usually in the range of 0.1 to 0.2 mg/ml of serum (about 12 μM). Affinity of the antisera for the antigen is determined by preparing competitive binding curves, as described, for example, by Fisher, D., (1980).

Antibody preparations prepared according to either the monoclonal or the polyclonal protocol are useful in quantitative immunoassays which determine concentrations of antigen-bearing substances in biological samples; they are also used semi-quantitatively or qualitatively to identify the presence of antigen in a biological sample. The antibodies may also be used in therapeutic compositions for killing cells expressing the protein or reducing the levels of the protein in the body.

While the preferred embodiment of the invention has been illustrated and described, it will be appreciated that various changes can be made therein by the one skilled in the art without departing from the spirit and scope of the invention.

References

Altschul et al., 1990, J. Mol. Biol. 215(3):403-410 / Altschul et al., 1993, Nature Genetics 3:266-272 / Altschul et al., 1997, Nuc. Acids Res. 25:3389-3402 / Beaucage et al., *Tetrahedron Lett* 1981, **22**: 1859-1862 / Berthon P. et al., 1998, Am. J. Hum. Genet., **62** : 1416-1424. / Brown

EL, et al., *Methods Enzymol* 1979;**68**:109-151 / Chai H. et al., 1993, Biotechnol. Appl. Biochem., **18**:259-273 / Chee et al., 1996, Science, **274**:610-614. / Chen and Kwok *Nucleic Acids Research* 25:347-353 1997 / Chen et al. (1987) *Mol. Cell. Biol.* 7:2745-2752. / Compton J. (1991) *Nature*. 350(6313):91-92. / Davis L.G., et al., Basic Methods in Molecular Biology, ed., Elsevier Press, NY, 1986 / Engvall, E., Meth. Enzymol. 70:419 (1980) / Feldman and Steg, 1996, Medecine/Sciences, synthese, **12**:47-55 / Fisher, D., Chap. 42 in: Manual of Clinical Immunology, 2d Ed. (Rose and Friedman, Eds.) Amer. Soc. For Microbiol., Washington, D.C. (1980) / Flotte et al., 1992, Am. J. Respir. Cell Mol. Biol., **7** : 349-356. / Fodor et al. (1991) *Science* 251:767-777. / Fuller S.A. et al., 1996, Immunology in Current Protocols in Molecular Biology, Ausubel et al. Eds, John Wiley & Sons, Inc., USA / Gonnet et al., 1992, Science 256:1443-1445 / Green et al., *Ann. Rev. Biochem.* **55**:569-597 (1986) / Griffin et al. *Science* **245**:967-971 (1989) / Guatelli J C et al., Proc. Natl. Acad. Sci. USA, **35** : 273-286. / Hacia JG, et al., *Nat Genet* 1996;**14**(4):441-447 / Haff L. A. and Smirnov I. P. (1997) *Genome Research*, 7:378-388. / Harju L, et al., *Clin Chem* 1993;**39**(11Pt 1):2282-2287 / Harlow, E., and D. Lane. 1988. Antibodies A Laboratory Manual. Cold Spring Harbor Laboratory. pp. 53-242 / Henikoff and Henikoff, 1993, Proteins 17:49-61 / Higgins et al., 1996, Methods Enzymol. 266:383-402 / Hillier L. and Green P. *Methods Appl.*, 1991, **1**: 124-8. / Huang L et al., 1996, Cancer Res; **56**(5):1137-1141. / Huygen et al., 1996, Nature Medicine, **2**(8):893-898 / Izant JG, Weintraub H, *Cell* 1984 Apr;**36**(4):1007-15 / Julan et al., 1992, J. Gen. Virol., **73** : 3251 – 3255. / Karlin and Altschul, 1990, Proc. Natl. Acad. Sci. USA 87:2267-2268 / Koch Y., 1977, Biochem. Biophys. Res. Commun., Vol.74:488-491 / Kohler G. and Milstein C., 1975, Nature, **256** : 495. / Kozal MJ, et al., Nat Med 1996;2(7):753-759 / Landergren U et al., 1988, Science, **241** : 1077-1080. / Lenhard T. et al., 1996, Gene, **169**:187-190 / Lin Z, Floros J, 1998, Biotechniques, **24**(6):937-940 / Livak et al., *Nature Genetics*, 9:341-342, 1995 / Livak KJ and Hainer JW, 1994, Hum. Mutat., **3**(4) : 379-385. / Lockhart DJ, 1996, Nat Biotechnol, **14**(13):1675-1680 / Mackey K, et al., 1998, Mol Biotechnol, **9**(1):1-5 / Marshall R. L. et al. (1994) *PCR Methods and Applications.* 4:80-84. / McLaughlin et al., 1989, J. Virol., **62** : 1963 – 1973. / Muzyczka et al., 1992, Cuur. Topics in Micro. and Immunol., **158** : 97-129. / Narang SA, et al., *Methods Enzymol* 1979;**68**:90-98 / Neda et al., 1991, J. Biol. Chem., **266** : 14143 – 14146. / Nickerson D.A. et al. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87:8923-8927. / Nyren P, et al., *Anal Biochem* 1993;**208**(1):171-175 / Nyren P. et al., 1993, Anal. Biochem., **208**(1) : 171-175. / O'Reilly et al., 1992, Baculovirus expression vectors : a Laboratory Manual. W.H. Freeman and Co., New York / Ohno et al., 1994, Sciences, **265**:781-784 / Ouchterlony, O. et al., Chap. 19 in: Handbook of Experimental Immunology D. Wier (ed) Blackwell (1973) / Pastinen et al., *Genome Research* 1997; **7**:606-614 / PCR Methods and Applications" (1991, Cold Spring Harbor Laboratory Press / Pearson and Lipman, 1988, Proc. Natl. Acad. Sci. USA 85(8):2444-2448 / Pietu G, 1996, Genome Res, **6**(6):492-503 / Rossi et al., *Pharmacol. Ther.* **50**:245-254, (1991) / Roth J.A. et al., 1996, Nature Medicine, **2**(9):985-991 / Roux et al., 1989, Proc. Natl

Acad. Sci. USA, **86** : 9079 – 9083. / Sambrook, J. et al.. 1989. Molecular cloning: a laboratory manual. 2ed. Cold Spring Harbor Laboratory, Cold spring Harbor, New York. / Samson M et al., 1996, Nature, **382**(6593):722-725. / Samulski et al., 1989, J. Virol., **63** : 3822-3828. / Sanchez-Pescador R., 1988, J. Clin. Microbiol., **26**(10):1934-1938 / Schena et al., 1995, Science, **270** : 467-470. / Schena et al., 1996, Proc. Natl. Acad. Sci USA, **93** : 10614-10619. / Schwartz and Dayhoff, eds., 1978, Matrices for Detecting Distance Relationships: Atlas of Protein Sequence and Structure, Washington: National Biomedical Research Foundation / Sczakiel G. et al., 1995, Trends Microbiol., 1995, **3**(6):213-217 / Shoemaker DD, et al., *Nat Genet* 1996;**14**(4):450-456 / Smith et al., 1983, Mol. Cell. Biol., **3**:2156-2165. / Sosnowski RG, Tu E, Butler WF, O'Connell JP, Heller MJ, *Proc Natl Acad Sci U S A* 1997;**94**(4):1119-1123 / Syvanen AC, *Clin Chim Acta* 1994;**226**(2):225-236 / Tacson et al., 1996, Nature Medicine, **2**(8):888-892. / Thompson et al., 1994, Nucleic Acids Res. 22(2):4673-4680 / Tyagi et al. (1998) *Nature Biotechnology*. 16:49-53. / Urdea M.S., 1988, Nucleic Acids Research, **11**: 4937-4957 / Urdea MS et al., 1991, Nucleic Acids Symp Ser., **24**: 197-200. / Vaitukaitis, J. et al. J. Clin. Endocrinol. Metab. 33:988-991 (1971) / Vlasak R. et al., 1983, Eur. J. Biochem., **135**:123-126 / Wabiko et al., 1986, DNA, **5**(4):305-314 / Walker G T et al., 1992, Nucleic Acids research, **20** : 1691-1696. / White, M.B. et al., *Genomics* 1997; **12**:301-306.

SEQUENCE LISTING FREE TEXT

The following free text appears in the accompanying Sequence Listing :

Homology with sequence in ref

Polymorphic base insertion of

Complement

Diverging amino acid in ref

Artificial sequence

Sequencing oligonucleotide primer

<110> Genset SA

<120> A nucleic acid encoding a geranyl-geranyl-pyrophosphate synthase (GGPPS) and polymorphic markers associated with said nucleic acid.

<130> D.18362

<150> US 60/093,940
<151> 1998-07-23

<160> 11

<170> Patent.pm

<210> 1
<211> 17131
<212> DNA
<213> Homo sapiens

<220>
<221> exon
<222> 486..546
<223> exon 1

<220>
<221> exon
<222> 633..826
<223> exon 1bis

<220>
<221> exon
<222> 7292..7384
<223> exon 2

<220>
<221> exon
<222> 13760..13830
<223> exon 3

<220>
<221> exon

<222> 14063..15251
<223> exon 4

<220>
<221> misc_feature
<222> 486..546
<223> homology with sequence in ref embl : AA398854

<220>
<221> misc_feature
<222> 7292..7384
<223> homology with sequence in ref embl : AA398854

<220>
<221> misc_feature
<222> 13760..13830
<223> homology with sequence in ref embl : AA398854

<220>
<221> misc_feature
<222> 14063..14314
<223> homology with sequence in ref embl : AA398854

<220>
<221> misc_feature
<222> 633..826
<223> homology with sequence in ref embl : Z44596

<220>
<221> misc_feature
<222> 7292..7384
<223> homology with sequence in ref embl : Z44596

<220>
<221> misc_feature
<222> 13760..13830
<223> homology with sequence in ref embl : Z44596

<220>
<221> misc_feature
<222> 14243..14670
<223> homology with sequence in ref embl : AA435858

<220>
<221> misc_feature
<222> 15055..15251
<223> homology with sequence in ref embl : AA194600

<220>
<221> misc_binding
<222> 14036..14081
<223> 5-187-77

<220>
<221> allele
<222> 14058
<223> 5-187-77 polymorphic base insertion of T

<220>
<221> primer_bind
<222> 13982..14000
<223> 5-187.pu

<220>
<221> primer_bind
<222> 14390..14409
<223> 5-187.rp complement

<220>
<221> misc_feature
<222> 1847..1848,6130,6145,10814,12943,13125,14874..14875,14917
15085..15086
<223> n=a, g, c or t

<400> 1
tcgggctccc tggttggggg gaggggacg acgaaaaatc ccccccggac tggaggtccg 60
ggcccccaat cgcgctgccc tccagaggac ggcggcgatg gaccctctgc agctccctcc 120
gggcaaaggt ccaggcggtg gccgtggcgg cggcaagatg aagctcaaga gtctccctcc 180
gcttcggcga ccgagctcct cactccggac tcgactgacg ggcaaacatc gcttcccccc 240
caccgactct aggttccccc cctttctccc ctcccctaga ttttttttcc cccctcccc 300
tacctctttc ccggatggcc tcttagacga ccttggattg gttaaagttc tttagaaccc 360
gcctatacac tgttcctatt ggtccctgga tacaaacaac gacgccattt tcccaccagt 420
tctatggaaa cagaaagtta cgcctcaagg ctttctggga aataaagtcc atactctggg 480
gccaacgcgc aaatcctcgt ccgcgagaac tgcaaggccc gcaatgccct gcgcctgcgt 540

ggaccggtgc gggggcgggg gggaggtgaa aggggcgggg caacaaagca gtagggaggc 600
ggcaacgacg cctgcgcagt gtgaccggga tggcgcattt tcttgcacca actaatgcgg 660
tgtcgctggc ggctgaggag ggcggagagt tctgtggtga aatagtggga aggattcatg 720
taggcatcgg gaagagccta agtccacatt ataaaatagg aagttgatgc ggggtacagt 780
tactcccgga ccggcggcgt gaaagtcgtg atatcatcgt tgaactgtga gcggcagtgg 840
cggcggctgg ggggaacccg gatgggaaga agggcggggg aggctgggag gcggggcaga 900
ggaaagaaag aaaggagagt gaggacccgg atgctgaacc ggattgtgta tgaattttcc 960
atcccctagc tttaagcgag gagggagagg aagggttggc caagtggggc ggaagggagc 1020
atctgagcga ggaggaagca gaaacctcac cgtttcttcc cctccggact ctgtgctagc 1080
actgtatacg tttgcagttc tctgcccagc cgctgtggaa aatcggcctc gaagtgattg 1140
aaattccctg tttatatcag gcggcttctt tcagatccat cgtctttctc ccggagtatg 1200
aatggaagga ttcagtatgc gcttcacatt tgtatgtctc tggccattct caaaccaggc 1260
ccttcccttt gaaaagtctt ttgcatggga tgttcacttc ttagacgcaa ggttgtgtgc 1320
cctggtttca tcgtctaacg cgttagaagg cgctttcatt tcttcatggg tgttgagcgc 1380
cgaccactgg ggtggcctct gccttcgtag acctgcgcct ggtgagacgg acagatgctg 1440
aacaaaacga tgtgaaatta ccgcagtggc agtgccccag aggagagttc cacggtgata 1500
ggagaatgag ggaatttggc ttctttaggg agggaaagga agggtttctg agcaagtgag 1560
gatcgagctg agagctgaag ggctagcagg agttaactaa ggaaagagaa aaggaaaaga 1620
cattccagac aaaaaggcta acttgtcaga aagccctgtg gcggaaggga gcttttccaa 1680
tatgaagaac tgagcctgga gagatgggat gagggggagt gtcgaacctt ttaggctttg 1740
taaaggagtt ttggttttct cctaatagca atgggatatc ttccaaggaa tctcaatcaa 1800
aagggagaga tggctccgat tggaatgtca tccctggctg aagagtnnag gaagcgaaaa 1860
aaagaagagt taaagaggca aatgcaggga acccgacgag gaggctattg ccgtagtagt 1920
tcacatggtg aaaagaatgg agcgtttgta ttaatgatta tggattcact ctttgaacaa 1980
atttctggca gctttttagt tttgaaagtg agaagtttca gactctcact gaggtattct 2040
gtagtttttt cactctaaaa ggaaactagt agagttcatg taacacacac taatgcctct 2100
ttacatttaa ctttagtatg tgatagctga aatttccagc tgtgataaat tgggaaatcc 2160
tttgatttaa aagaaaaaca aaggcgggtg agggtgagag tatatgccac ggtgtgtaga 2220
atcctttaga ctcttaagaa gacacaaggc ggctgggcgt ggtggctcac gcttgtaatc 2280
ccagcacttt gggaggccga ggcgggcgga tcacgaggtc aggagatcga gaccatcctg 2340
gctaacacgg tgaaagcccg tctctactaa aaatacaaaa aaattagccg ggcaaggtgg 2400
cgggcgcctg tagtcccagc tactcgggag gctgaggcag gagaatggcg tgaacccggg 2460
aggcggagtt tgcagtgaga cgagatcacg ccactgcact ccagcctggg cgacagagtg 2520
agacgctgtt tcagaagaaa gacacaaggc aagttggttg tcgatacctg gaaaaattga 2580
agttcttatg ttttcatacc actgaaaatg cttgtatgta aatatcctct gggacaggaa 2640
attgacttaa gtgagtattc ttaaacatct ctaagtgagg aaaggaaata ttttttaaag 2700
cataattagt gttttaagtt gaaaaataac atcaaccaca aagctctacg aattgaaaca 2760
aagattagct ctgatttctg tgcaacaggg tacacctgtt acaggtcctg acacaaaagg 2820
gaattctgaa agtgcatctc attgattttt aagttcggtc aaatgtgttt tggaggctgt 2880
gagaaaatat acaaacgtga ttcttgctcc caacttgtag ttgagaaaag atagatacta 2940
acatttaaat agagaagtat atgagatcct tttttaattc tacttttaat gatgttcgat 3000
aataatcttt tagctaagcc attattcttc ctgttttgca tcttcttttc ttacttcaat 3060

ccctgataat aaggtcacgt gtcagagatc aaatagtata ggtaataggt tacctaaata 3120
ggtatttgca taataggtta cctaactaaa taggtttttg cctaataggt atgttgatta 3180
tttcgcttac ttgattcttt atgagccttt ttttccttgc gacgtctttg gtattaattg 3240
ttagtcaaga tggatgtaga aatttttccat atgggatgtt tctctttgaa ttcatgttgt 3300
taaaatgatt tcttttggtg gagtgctgat ctttttttatg attgtttcat atagataaga 3360
acagactaca aaaaaatatg cctttcaatc ctgaagagta acctgaacta tacactagtt 3420
ttgtgcttta attttcattt gtaatctgcc ttcaataaag agttaagcta gtggaattta 3480
tgtcttagct tgttataaca caaacacgaa tatttgtctg cttggcatta aagggtaaag 3540
atattccata gctgggaatc ttaatctgag gtacgtgtaa acattcaggg actatatgat 3600
ctctgagaat ttgtatgttg taagtctttg tggcagtgta tacatttgtg ttgcaactta 3660
ttaacacata caccgggctt ttttttttt ttttagaaga ttcatagctt tcatcatatt 3720
ctcaaaaggt ttctgtgacc catgagatgg tttacagtat ggggaagcat caaagcactt 3780
gcacagttga tggttatatg tgtgtgttat tatttcagcc acccattatc atgtgcttac 3840
caactgccta acagtgcata catatgtaga agttttattc ttttctcctg ttgccatatt 3900
atacgtctca tttcacagca gaaaaacaac tgcatgacag agacaatgtg gttcaaacca 3960
ttttaccctt gtattcattg actgctacaa aacaggaaca ttaaatacct gattgtcacc 4020
aaattgggta gtctcagcac ttctacactc gtaattgtgc tggaaaagtg gaatgctagc 4080
actaataatt agattttggt ttggagggtt ttttatttgt ttattcttac ttgtataaat 4140
ttatggggtg caagtgtagt tttatcacat gcatagattg cattgtagtg aagtcaggac 4200
ttttaggggg tccatcaccc atgtaatcac gttgtaccca ttaagtaatc tttcatcatc 4260
cacctccttc ccaccttctc acccttttgga atctccattg tctatcattc cacactccat 4320
gtccatgtat acacattatc tagctcccat ttataattga gaagatgtac tatttgtctt 4380
ttatgtctga cttgttacac ttaaggtaag ggctatccat ccattttgct gcaaatgaca 4440
tgatttcatt ttgttttaat ggctgagtaa tcattcgttg tatatatacc acattttctt 4500
tattcagtca tctgctgatg gacacttagg ttgattccat atctttacta ttgtgaatag 4560
tgctgtaata aacacatagt gcaagatttt ggaaatttta cttttgtggc acgttgttgg 4620
tatttactca ggatctttgg atttgcttgg ctgcatgtat atgaatcagt gtgtttattt 4680
actgaaatat gtgcaaaagt cttgtctttg gtggattaat ttataatata aatccacaaa 4740
agtcagattc tgctcctaag tatattttac atttttaaat ttaatgccag caagaagtta 4800
cagtactaga attgccttac ccctgagagt atcaatgatc agatcatagt atcaggtgac 4860
tgggctatag aagatgactt ttattactta acattatgaa gttactaggg ctgatttaga 4920
aatcgaggaa cactggtgaa accccgtctc tactaaaata caaaaattag ctgggcgtgg 4980
tggtgggcac ctgtagtccc agctactcag aaggctgagt caggagaatt gcttgagccc 5040
aggaggcaga ggttgcagtg agccgagatc gtgccactgc actccagcct gggcgacaga 5100
gtgagactcc gtctcaaaaa aaaaaaaaaa aaaaaaaaag gaacacatcc tcactgttac 5160
aataaataac agtagcccac acccccttag ttgtgatgtg gtgtgatacc atgtaagcaa 5220
cctatttcca gttcccctaa cattctcaag cagctgtatc agaatcatac aagatgcata 5280
tttaaattga agatttctaa gtctctggcc cagacttaga aaaaaaggat caggccgggc 5340
acagtagcta acacctgcaa ttccaacact ttgggaggct gaggcgggtg gatcgcctga 5400
ggtcaggagt tttgagacca gcctggccaa catagtgaaa ccccatctct actaaaaatt 5460
caaaaaatta gctgggcgtg gtggcaagaa cctgtaatcc ctgctattcg ggaggctgag 5520
gcaggggaat cacttgaacc cgggaggtgg aggttgcagt gagccaagat tgcgccactg 5580

cactccagcc tgggcaacga gcaaaactcc gtctcaaaaa aaaaaaacaa aaggaccttt 5640
gagcaatcag aataacacaa agtacatgaa ctgaacttca ttttcttcat tcaaaagaaa 5700
gtggccctca ctcaagcaaa tatattcttg tgctttatct tctggcatac tgagataact 5760
ttctaaagtg gtttccaatt ccaaaatcca atgatgtgca actcattgaa cagccctaac 5820
cacaaactgc cattagatgc catattacat ttagcctttt tgttgtagaa aagttggtta 5880
gaagtgggct caggattcta aagactaaat catagtccca agaagcaaaa gaaagaggat 5940
aaaagtaata aacttcccaa aatgtgccaa agatgctaga gcagttagat tcctaatatg 6000
aggacaagta ataatagaaa cagatacaaa gaaataaagt agagattcaa cagtacaggg 6060
agaccctagg aagaccatga gtgttattct aggaaatact gaaataagac agatttcagt 6120
ataaaggggn aatatgttta ataanatata tgcatttgag ttaatgcgta ttttaaatca 6180
gaaatctctg aaatggattg attgtagaga aactactagg gggacgagga gaatcccttt 6240
aaattttaaa tacataaaac atactcatct tagtgctcat ttaaaaaagg atatgtttac 6300
taattagtgt aatcagttaa atacagaggt atctttccaa ttctttggat gtgttttgac 6360
atttgccgtc aacaaattaa gccttttgtg gttgattaaa ataggaaaag cttaatataa 6420
gttatgtgac taagaaaaca acttaaaaac caagacaaca ctttgaccaa tataatcact 6480
tgaatgaaga attttctaat tgagatataa tttacatacc acccatttaa agtgtacatt 6540
tcagcagttt ttagtgtatt cacagggctg tgcaaccatc acaatttaat tttataacat 6600
tttgatccct gcgaaaagaa accctgtact cattagcaat tagtccctgt tcctaaccac 6660
taatctactt tctttctctg tagattggct tattctgaac atttcgtata aatggaatca 6720
tacaatatgt agtctcttga gattggcttc tttcacttaa catgttttca aggcttcata 6780
gctgtagaat cttgctttgt ttttttgaga ctggagtcac tctttcgccc aggctggagt 6840
gcagtggtgt gatctcagct cactgcaacc tctgcctccc gggttcaagc agttctcctg 6900
cctcagcctc ccaagtagcc agaactacag gcacacacca ccatgctcgg ctaatctttg 6960
tagttttagt agagatggtg tgaaggctgg tctcgaactc ctgacctcat gatctaccca 7020
cctcagctaa tttttcatat ttttagtaga gacaaggttt tgccatgttg cccaggctgg 7080
tctcgaactc ctgggcttaa gctatccgcc cgcctcagcc tcccaaagtg ctgggattac 7140
aggcgtgaac taccgtgccc agcaacagaa tcttcttttt aaaccagact aggtgtcttt 7200
tcacaaacac cctgcaatac aaattccttt gcagtttgac actgaaagat gattagtttc 7260
atgtgatctt tatgtttctc ctttttgaca gattagcttt gaagtttaaa tccaatggag 7320
aagactcaag aaacagtcca aagaattctt ctagaaccct ataaatactt acttcagtta 7380
ccaggtaata cttcacttac agtccatata gggtcatttt catgcagtag tggtcgttca 7440
aatgttagca aatagaaaag gttagacttg ctagccgttg agattttcta tttaaggtga 7500
tgcgtatgag aaaaatgata aatagaacat tataattttt tctttattaa aaggtaattt 7560
ttgccaggtg cagtgataca tacctgttgt cccacctact tgggaggctg aggcaggagg 7620
atggcttgag cccaggagtt taaggctata gtgcacaatg atcacacctg tgaatagcca 7680
ctacactcca gcttgggcaa catagtgaga ccccgtctct taaaaagaaa cgtaattttt 7740
gaaggcaccc tttaaaacat atccaattat ttaacatatc ttgaaaaata aaaatactta 7800
aaacattttg gtatctcatt ggaggttgta ctctttacgg atattacgca ttcagattcc 7860
ccactgttta gatattaggg gaagttacgc agatttgttt aacagtagaa cactttattt 7920
accatacatg ttcaagttta ccttctatgt ctgtattttc cagtatctca cacatacact 7980
gcatttcata tactactggt tcctttgaga gccaaataat aatgtatcta aaatcacagt 8040
atttggaaat atagcccact ttattcctgt ataagggtat gccaccttgg acatggcttc 8100

ctacctcacg tgtacgtgtg tgtttttgtt ttattttgct tctttaaaaa cttgtctgga 8160
ggctgggcgt ggtggctcac gcctgtaatc ccagcacttt ttgaggccaa ggcgggcgga 8220
tcatgaggtc aagaggttga gaccagcctg gccaacatgg taaaaccccg tctctactaa 8280
aaacacaaaa gttagctggg catggtggcg catgcctgta gtcccagcta ctcgggaggc 8340
tgaggcagga gaatcacctg aacctggaag gcagaggttg cagtgagctg agattgcatc 8400
actgcactcc agcctggcaa cagaatgaga ctccgactca aaaaaaaaga agaacttgtc 8460
tggaaatgat aataagcaaa aactcatgaa tataataaac aggggttatt gtaataaaaa 8520
atcatttgta ttagaatatt ctttctcata gacataatat aggccaggtg tggtggccca 8580
cacctgtatt cccagcactt tgggaagcca aggcaggatt gcttgagacc aaaagtttga 8640
gaccaccttg ggcaacataa caagtccccc tctctgtttt aaacattttt taaaaaagaa 8700
gaaataatat aaaagttggt aaattatttg acaagcataa aaacctattt agccatactg 8760
tgactaaact ctaatgatgc tctcaattca gtctcaatag acacttttaa atttccgtgc 8820
taaagtacac acctttcttt atgagcactt ctctgtggta atatgtgcat ttctgttctt 8880
catgagcctg ggaaggataa aagccaaaag aatgcttgct cctgtgctac accttggaaa 8940
ccataattag tgtcattttt attttggccg accctaatag agactcgcct gctaatgtca 9000
atgcatgaga agaatgaggg aatgacagaa atggagaatt caaagggaag gttgcccact 9060
gtttaagaaa aagccaagag actgcttttg agtgacattt atccagcagt tagtaactta 9120
tttcagtatc tcccagtgag aaacatggca cagtttcact ttcactctac ccagctctta 9180
ctgccagaca tcctttagaa cacgctcaca aacactagct ggaactgggc tggcattaat 9240
agcaagccag ttatcagtgc tgacaaaagt ctaacaagca tcgcttgaat gtctcttact 9300
ctgctactta caaagcaagg actgcctaca gttacatttt aaccataatg cttacttatg 9360
ctgtgaccac cttctgtgac ttccttttt ttaattctca ttacttggaa ataatgtttt 9420
aagacattag ataacatatt taaaattatc actaggtacc tcacctttt attcaagtac 9480
gttcttgatc catgatggaa tacaacctca aaagatacta ctaaagaaat atgacattgc 9540
actatgcaca taacacactt attttttac agagagcttc agagttacta aagtaactta 9600
gaggtgtgcc aggtcattta tactgttgta atattactct tgctaataaa taataataat 9660
gctatcagta ttttctgaag tcaacctggc caacatggtg aaaccctgta tctactaaaa 9720
atacaaatat tagccaagta tggtagcgca tgcctgtagt cccagctgag gcacgggagt 9780
cacaggagcc taggaggcag aggttgcagt gagccgagat cacgccactg cactccagcc 9840
tgggcaacag agtgagacac tgtctcaaaa aaaaaaaagg attttctgaa attagtaaag 9900
aaaattattt ttatttttaa atttctcata cttgctgtca tcttatgttt atgtttgttt 9960
atttgcctta gtgtggggcc ctagatgagg tgaagggtgg gattagggag agatgaagct 10020
ggcagtggag gaagaagggc tccaaaaaga gagacaataa tgtttagatc ttaaagagga 10080
agcagtaatc ttttaatttt gagagatctc tgtgattagc ctcagtacta gaaattattt 10140
tggaactcag ccaggcgcgg tggctcacat ctgtactccc agcactttgg gagaccgaag 10200
tgggcagatg gcttaagccc aggagttcaa gaccagcctg ggcaacatgg caaaaccctg 10260
tctctactaa aaatacaaaa aattagccag gcatgtgata cgcccttgta gtcccagctt 10320
acctgggggа ctgaggtggg atgattaccg gagcctggga ggttgaggct gcagtaagcc 10380
aagatcacac cactgcaccc cagcctgggt gattaaggga gaccccgtct cagaaaaaaa 10440
aaaggggggg aaacttaaaa gcatcaggct aaacactagc atgtcatcag aggggaaaaa 10500
aatattaaaa ctgtagtacc tcaaaaataa gccatatatt gtactgtttt ctatataaca 10560
ttcaaaagta aaatgaaaaa tgaaatttca cattgagact ctgtttttca tcttcaaaaa 10620

aatgtgttta agtgatacag gccaagtgca gtggctgact tattatccca gcactttggg 10680
aggccaagtg ggacagattg cttttgagcc caggggtttg agaccagcct gggcaacagg 10740
gcgaaaccct gcctctacaa aaaataaata aataaaaata aaattagcca ggcatggtgg 10800
cttgttcttg tagntcccag ctactcaggg gacttgagcc taggaggtca aggctgcagt 10860
aggccgtgat tgtgccactg cactccagcc tgggtgacag agcgagaccc tgtctcaaaa 10920
ataataataa taggccgggc gtggtgggtc acacctgtaa tcccagcact tcgagaggcc 10980
aaagcatgtg gacgacttga ggtcaggagt tcgagaccag cctggccaac atggggaaac 11040
cctgtctcta ttaaaagtac aaaaaattgg ccgggcgcgg tagctcacgc atgtaatccc 11100
tacactttgg gaggctgagg tgggtggatc acctgaggtc aggaattcaa gaccagcctg 11160
gccaacatga tgaaaccgtc tctactaaaa atacaaaaaa ttagctggat ttagtggcgc 11220
acgactgtaa tcccagctac tcaggaggct gaggcaggag aatcgcttga acctaggagg 11280
tggaggttgc agtgagccaa gatcgtgaca ctgtacccca gcctgggcaa caagagcaaa 11340
actcgatctc agaaaaaaaa tacaaaaaat tagctaggcg tagtgacgca cacctgtaat 11400
cccagctact cgggaggctg agacaggaga atcccttgaa cccaggaggc gaaggttgtg 11460
gtgagccgag ccaagatcgt gccattgctt tccagcctag gtgacagagc aaaacttcat 11520
ctccacaaac aaacaaacaa acaaaaaaac ccataatccc agcattttgg gaggccaaca 11580
caggtgaatt acctgaggtc aggagtttga caccagcctg gccaacatag tgaaaccctg 11640
tctctactaa aattacaaaa attagccagg tgtggtggca ggtgcctgta atcccagcta 11700
cttgggaggc tgaggcagga gaatcgcttg aacccagggg gcggaggttg cagtgagccg 11760
agatcacacc attgcactct agcctgggtg acaagagcga aattccatct ccaaaaaaaa 11820
aaaaagaaaa cagtatttta gttttaactt tttatgtaac cattttcctg aaaccttatc 11880
taaaattagg atgttattac catgcattca tttagcagaa aacttataga acatttttac 11940
taagtgaact ggccatggtt tttatctatc attcctttgt atgtgactac aatgacttct 12000
agtggtaact tctatccaaa gacctatctt aaattagcca ggcatggtgg cacatgcgtg 12060
taatcccagc tactcaggag gctgaggcag gagaatagct tgatcttggg aggcggaggt 12120
tgcaagtgag ccgagatcac gccgctgcaa tccagcctgg gcaacagaat gagactccgt 12180
ctcaaaaaca aaaaacaaaa agacctatct tgagctttcc gtgtaagaaa aagatgatac 12240
tgttgggtga agtgactcaa cgtctgtaat ttcagcaatt tgggaggctg tagcggccgg 12300
attgcttgag cccaggagtt tgagaccagc ttgggcaaca tgggaacaca ctgtctctac 12360
aaaaacaaaa attaaccggg cgtggtcgct tgcacctata gtgccagcta ctcgggaggc 12420
tgaggtggag gctgcagtga gctgtgaaca caccactgca ctccagcctg ggtgacagag 12480
tgagaccctg tctcaaaaaa aaaagcaaga agcgcagtgg ctcacgcctg taatcccagc 12540
actttgggag gccgaggcgg gcggatcacg aggtcaggag atcgagacca tcctggctaa 12600
cacggtgaaa ccccgtctct actaaaaata caaaaaatga gccgggcgtg gtagcgggcg 12660
cctgtagtcc cagctactcg ggaggctgag gcaggagaat ggcgtgaacc cgggaggcgg 12720
agcttgcagt gagccgagat cgcgccactg cactccagcc tgggcgacag agcgagactc 12780
cgtctcaaaa aaaaaaaaaa aaaaaaaaac aagaaagaaa aaaagaagat actgaaaaat 12840
agatgtccct agtcaaaata atgagattag cttttgacta aactcaggat attaaaaggg 12900
aatacttcag tgcatgatga tctcattttt gaaaggaaag aancagagct tccccatctc 12960
taaaacctta attcaaagga gaaatagata atttcaagag gtatttttat gaggtaatag 13020
taaaatatat tttattaaca gtacctatag ttatgtaaaa taggtagtgc caattaactg 13080
acactaaact agcttcttgg cctggcgcag tggctcacgc ctgtnaatcc aaacactttg 13140

ggaggccgat gcgggtgtat cgcttgggct caggaattca aggccagcct gggcaacata 13200
ttaaaacccc ctttctataa aatatacaaa aattagccag gcatggtgtg tgcctgtagt 13260
cccagatact caggaggctg aggcacgaga atcatgtgaa cccaggaggt ggagtttgca 13320
gtgagccgag atcacgccac tgcactccag cctgggcaac agagcaaaac tctgtctcaa 13380
ataattaata aataaactag cttccttttc aaaaaaagaa ataaattagg tcctaagtcc 13440
taaaagccca tcctacttta aaattgttta ttcaagttca gatgaaaaga gtggactagt 13500
aggcaactga agtgctttag agtctcccgt gcctgcccta attttagaag gttgtgcact 13560
ttatgatcca gatttctgag tggttgagaa tgagttattg agcagtgcaa ggcaagctct 13620
gcagtaggta atggattgat gaggctggat ttagcaagtc tgatcaatct aaaggaagtt 13680
tctgaatgtg ttttttgtag ttaaaatact cataattaaa acacttatca cattgtcaca 13740
ttttattttt aaattgcagg taaacaagtg agaaccaaac tttcacaggc atttaatcat 13800
tggctgaaag ttccagagga caagctacag gtattaggca actctaacct cattaatccc 13860
caagaaatta atagctgtcg cataaaaata ttcctagttc ttgattgaat ttagtcctca 13920
tgcaagatat tattttatat tgaggttgct aaatatttat tagttgtgaa aattaacaca 13980
cctgagactt tcataatctg ttaattaaac tgagtaagtt ttgaatagtt caaataagtg 14040
aaattttcaa tttttttatt agattattat tgaagtgaca gaaatgttgc ataatgccag 14100
tttactcatc gatgatattg aagacaactc aaaactccga cgtggctttc cagtggccca 14160
cagcatctat ggaatcccat ctgtcatcaa ttctgccaat tacgtgtatt tccttggctt 14220
ggagaaagtc ttaacccttg atcacccaga tgcagtgaag ctttttaccc gccagctttt 14280
ggaactccat cagggacaag gcctagatat ttactggagg gataattaca cttgtcccac 14340
tgaagaagaa tataaagcta tggtgctgca gaaaacaggt ggactgtttg gattagcagt 14400
aggtctcatg cagttgttct ctgattacaa agaagattta aaaccgctac ttaatacact 14460
tgggctcttt ttccaaatta gggatgatta tgctaatcta cactccaaag aatatagtga 14520
aaacaaaagt ttttgtgaag atctgacaga gggaaagttc tcatttccta ctattcatgc 14580
tatttggtca aggcctgaaa gcacccaggt gcagaatatc ttgcgccaga gaacagaaaa 14640
catagatata aaaaaatact gtgtacatta tcttgaggat gtaggttctt ttgaatacac 14700
tcgtaatacc cttaaagagc ttgaagctaa agcctataaa cagattgatg cacgtggtgg 14760
gaaccctgag ctagtagcct tagtaaaaca cttaagtaag atgttcaaag aagaaaatga 14820
ataatgttaa gccattcttg attggacctc atagcttatt ttagttaatc tttnntttgt 14880
cttttagcct taccaccttt taaaaaattt gttattntcc agaaacagta aataggtgag 14940
taggggtggt gcaagtgaat tcgttttcat ttagaagccc ctctgtacag ataatcaaaa 15000
ttcaaagttg aaagaatcaa aagcagccac agttatgtag gtctgatttg aatgtcataa 15060
ttgcagtgac aggacattgc caccnnctcg tatcctacta ccatcaatgt tgtgtttatt 15120
ccgtcaataa aaaagacttg cttccaggaa tttttatcca tacactttct aactgtacta 15180
tctgggcagt tccaagccag tttctattag ctagctggac caaagaccac aaatctcttt 15240
ttttcctaaa cgctgctgta aggaatatct cacttttccc cccggaaaca ccctcactga 15300
agtcttctat gaaaaggcct gataatgggc tgggcgcggt ggctcacgcc tgtaatccca 15360
gcactttggg aggccgaggc gggcagatca cgaggtcagg agatcgagac catcctgaca 15420
cggtgaaacc ctgtctctac taaaaataca aaaaattagc tgggcgtggt ggtgggcgcc 15480
tgtagtccca gctactcggg aggctgaggc aggagaatgg tgtgaaccca ggaggcggag 15540
cttgcagtga gccgagatag tgcctctgca ctccagcctg ggtgacagag cgagactccg 15600
tctcaaaaaa aagggctgat aatgataaac agtgagcact ccggtccttt ttcttaggtt 15660

ttccttttt ccttcctctc caccccacaa gttttgcttt ttaaccaagg tgtctctgct 15720
tgatgaaatt cacatgctag tctaaatctt tttttctccc ttgtaacatt tatgtgcccc 15780
aaactggtta gtatatgggt acagcattcc ctttccaatt gggaagcgga aaaagagagt 15840
atgggatatt ttagaaggga gcctttgaac cttattatat ttccccatca ttgatagtga 15900
caatcttaaa agggttgttt tcttacctta agtacaaaag catggaaaaa tgcgcttttc 15960
cttcccgccc acatcaccac cccgacttga agacagtagg tgcttgaatg gaaagtgagt 16020
aggcatcttt aatcgccctg attaaaggaa agtgttagcc tgagagggcc tgactgaaaa 16080
gtaaccaaag gcttaatatc aaacactaat tagcttttta gtgccttaac cctgacctgg 16140
ttaccagttt tctgtagttt ctacacccaa gccactgaag tcatctgtgg cccaagaggt 16200
aggacaaaaa aaaaaaaaaa aaaaaagctg atttcaatat ttgatttgtt gacatcccaa 16260
aatgaaagtt ttatgtttcc cttagaaaca tgttttgctt ggttctatag tatgttactt 16320
aggatctatt taccatatat ttgtatgaga aatcctcacc caagcattca acctaaatct 16380
ttgaaaagtt gggtgctgtc tttagtaact tttaaaatag tttaaatctc ccattttaat 16440
agtgataagg aaacctgtta aaatcatggc tattgatgtt atagtatgga aagttgaact 16500
ttatgaaccc atacttttaa aaagcatttt taaaaatcta acactgacta tagaaacaaa 16560
ttaaaatgtc tacctttaag tataaaaatt gcttaagtag atttgttcct tgcctatcaa 16620
attaattttg gcctggtgtt cttcattatt catttgttaa ttttatcttg cctttgtcaa 16680
taacagaaat gtttgtcatt gaattgggaa ttttttttt ttttttgag acggagtttc 16740
actcttgttg cccaggctgg agtgcaatgg cgtgatctca gctcactgca acctccacct 16800
cccgggttca agcgattctc ctgcctcagc ctcctaagta gctgggatta cagatgcctg 16860
ccatgttgcc tggctaattt ttttttttt ttttttttta agtagagatg gggtttcacc 16920
atgttggcca ggctggtgtt gaacttctga cctcaggtga tccagctgcc tcggcctccc 16980
aaagtactgg gattacaggc atgagccacc gcacccagcc aaattgggga cttttaacag 17040
tcattttacc tgtagaataa tcaaaactct tcacttgatc tgtagtcata gctattaaca 17100
cagaaaaatg aatgccagtt atgttgccat a 17131

<210> 2
<211> 1414
<212> DNA
<213> Homo sapiens

<220>
<221> CDS
<222> 85..987

<220>
<221> polyA_signal
<222> 1289..1294
<223> AATAAA

<220>
<221> misc_feature

<222> 1..477
<223> homology with sequence in ref embl : AA398854

<220>
<221> misc_feature
<222> 406..833
<223> homology with sequence in ref embl : AA435858

<220>
<221> misc_feature
<222> 1218..1414
<223> homology with sequence in ref embl : AA194600

<220>
<221> misc_feature
<222> 1037..1038,1080,1248..1249
<223> n=a, g, c or t

<400> 2

```
cgcgcaaatc ctcgtccgcg agaactgcaa ggcccgcaat gccctgcgcc tgcgtggacc      60
gattagcttt gaagtttaaa tcca atg gag aag act caa gaa aca gtc caa          111
                          Met Glu Lys Thr Gln Glu Thr Val Gln
                          1               5
aga att ctt cta gaa ccc tat aaa tac tta ctt cag tta cca ggt aaa        159
Arg Ile Leu Leu Glu Pro Tyr Lys Tyr Leu Leu Gln Leu Pro Gly Lys
10                  15                  20                  25
caa gtg aga acc aaa ctt tca cag gca ttt aat cat tgg ctg aaa gtt        207
Gln Val Arg Thr Lys Leu Ser Gln Ala Phe Asn His Trp Leu Lys Val
                30                  35                  40
cca gag gac aag cta cag att att att gaa gtg aca gaa atg ttg cat        255
Pro Glu Asp Lys Leu Gln Ile Ile Ile Glu Val Thr Glu Met Leu His
            45                  50                  55
aat gcc agt tta ctc atc gat gat att gaa gac aac tca aaa ctc cga        303
Asn Ala Ser Leu Leu Ile Asp Asp Ile Glu Asp Asn Ser Lys Leu Arg
        60                  65                  70
cgt ggc ttt cca gtg gcc cac agc atc tat gga atc cca tct gtc atc        351
Arg Gly Phe Pro Val Ala His Ser Ile Tyr Gly Ile Pro Ser Val Ile
    75                  80                  85
aat tct gcc aat tac gtg tat ttc ctt ggc ttg gag aaa gtc tta acc        399
Asn Ser Ala Asn Tyr Val Tyr Phe Leu Gly Leu Glu Lys Val Leu Thr
90                  95                  100                 105
ctt gat cac cca gat gca gtg aag ctt ttt acc cgc cag ctt ttg gaa        447
```

```
Leu Asp His Pro Asp Ala Val Lys Leu Phe Thr Arg Gln Leu Leu Glu
                110                 115                 120
ctc cat cag gga caa ggc cta gat att tac tgg agg gat aat tac act      495
Leu His Gln Gly Gln Gly Leu Asp Ile Tyr Trp Arg Asp Asn Tyr Thr
            125                 130                 135
tgt ccc act gaa gaa gaa tat aaa gct atg gtg ctg cag aaa aca ggt      543
Cys Pro Thr Glu Glu Glu Tyr Lys Ala Met Val Leu Gln Lys Thr Gly
        140                 145                 150
gga ctg ttt gga tta gca gta ggt ctc atg cag ttg ttc tct gat tac      591
Gly Leu Phe Gly Leu Ala Val Gly Leu Met Gln Leu Phe Ser Asp Tyr
    155                 160                 165
aaa gaa gat tta aaa ccg cta ctt aat aca ctt ggg ctc ttt ttc caa      639
Lys Glu Asp Leu Lys Pro Leu Leu Asn Thr Leu Gly Leu Phe Phe Gln
170                 175                 180                 185
att agg gat gat tat gct aat cta cac tcc aaa gaa tat agt gaa aac      687
Ile Arg Asp Asp Tyr Ala Asn Leu His Ser Lys Glu Tyr Ser Glu Asn
                190                 195                 200
aaa agt ttt tgt gaa gat ctg aca gag gga aag ttc tca ttt cct act      735
Lys Ser Phe Cys Glu Asp Leu Thr Glu Gly Lys Phe Ser Phe Pro Thr
            205                 210                 215
att cat gct att tgg tca agg cct gaa agc acc cag gtg cag aat atc      783
Ile His Ala Ile Trp Ser Arg Pro Glu Ser Thr Gln Val Gln Asn Ile
        220                 225                 230
ttg cgc cag aga aca gaa aac ata gat ata aaa aaa tac tgt gta cat      831
Leu Arg Gln Arg Thr Glu Asn Ile Asp Ile Lys Lys Tyr Cys Val His
    235                 240                 245
tat ctt gag gat gta ggt tct ttt gaa tac act cgt aat acc ctt aaa      879
Tyr Leu Glu Asp Val Gly Ser Phe Glu Tyr Thr Arg Asn Thr Leu Lys
250                 255                 260                 265
gag ctt gaa gct aaa gcc tat aaa cag att gat gca cgt ggt ggg aac      927
Glu Leu Glu Ala Lys Ala Tyr Lys Gln Ile Asp Ala Arg Gly Gly Asn
                270                 275                 280
cct gag cta gta gcc tta gta aaa cac tta agt aag atg ttc aaa gaa      975
Pro Glu Leu Val Ala Leu Val Lys His Leu Ser Lys Met Phe Lys Glu
            285                 290                 295
gaa aat gaa taa tgttaagcca ttcttgattg gacctcatag cttattttag         1027
Glu Asn Glu  *
        300
ttaatctttn ntttgtcttt tagccttacc acctttaaa aaatttgtta ttntccagaa    1087
acagtaaata ggtgagtagg ggtggtgcaa gtgaattcgt tttcatttag aagcccctct    1147
gtacagataa tcaaaattca aagttgaaag aatcaaaagc agccacagtt atgtaggtct    1207
gatttgaatg tcataattgc agtgacagga cattgccacc nnctcgtatc ctactaccat    1267
```

caatgttgtg tttattccgt caataaaaaa gacttgcttc caggaatttt tatccataca 1327
ctttctaact gtactatctg ggcagttcca agccagtttc tattagctag ctggaccaaa 1387
gaccacaaat ctctttttttt cctaaac 1414

<210> 3
<211> 1547
<212> DNA
<213> Homo sapiens

<220>
<221> CDS
<222> 218..1120

<220>
<221> polyA_signal
<222> 1422..1427
<223> AATAAA

<220>
<221> misc_feature
<222> 1..359
<223> homology with sequence in ref embl : Z44596

<220>
<221> misc_feature
<222> 1170..1171,1213,1381..1382
<223> n=a, g, c or t

<400> 3
gcgcattttc ttgcaccaac taatgcggtg tcgctggcgg ctgaggaggg cggagagttc 60
tgtggtgaaa tagtgggaag gattcatgta ggcatcggga agagcctaag tccacattat 120
aaaataggaa gttgatgcgg ggtacagtta ctcccggacc ggcggcgtga aagtcgtgat 180
atcatcgttg aactattagc tttgaagttt aaatcca atg gag aag act caa gaa 235
Met Glu Lys Thr Gln Glu
1 5
aca gtc caa aga att ctt cta gaa ccc tat aaa tac tta ctt cag tta 283
Thr Val Gln Arg Ile Leu Leu Glu Pro Tyr Lys Tyr Leu Leu Gln Leu
10 15 20
cca ggt aaa caa gtg aga acc aaa ctt tca cag gca ttt aat cat tgg 331
Pro Gly Lys Gln Val Arg Thr Lys Leu Ser Gln Ala Phe Asn His Trp
25 30 35
ctg aaa gtt cca gag gac aag cta cag att att att gaa gtg aca gaa 379

```
Leu Lys Val Pro Glu Asp Lys Leu Gln Ile Ile Ile Glu Val Thr Glu
    40                  45                  50
atg ttg cat aat gcc agt tta ctc atc gat gat att gaa gac aac tca      427
Met Leu His Asn Ala Ser Leu Leu Ile Asp Asp Ile Glu Asp Asn Ser
55                  60                  65                  70
aaa ctc cga cgt ggc ttt cca gtg gcc cac agc atc tat gga atc cca      475
Lys Leu Arg Arg Gly Phe Pro Val Ala His Ser Ile Tyr Gly Ile Pro
                75                  80                  85
tct gtc atc aat tct gcc aat tac gtg tat ttc ctt ggc ttg gag aaa      523
Ser Val Ile Asn Ser Ala Asn Tyr Val Tyr Phe Leu Gly Leu Glu Lys
            90                  95                  100
gtc tta acc ctt gat cac cca gat gca gtg aag ctt ttt acc cgc cag      571
Val Leu Thr Leu Asp His Pro Asp Ala Val Lys Leu Phe Thr Arg Gln
        105                 110                 115
ctt ttg gaa ctc cat cag gga caa ggc cta gat att tac tgg agg gat      619
Leu Leu Glu Leu His Gln Gly Gln Gly Leu Asp Ile Tyr Trp Arg Asp
    120                 125                 130
aat tac act tgt ccc act gaa gaa gaa tat aaa gct atg gtg ctg cag      667
Asn Tyr Thr Cys Pro Thr Glu Glu Glu Tyr Lys Ala Met Val Leu Gln
135                 140                 145                 150
aaa aca ggt gga ctg ttt gga tta gca gta ggt ctc atg cag ttg ttc      715
Lys Thr Gly Gly Leu Phe Gly Leu Ala Val Gly Leu Met Gln Leu Phe
                155                 160                 165
tct gat tac aaa gaa gat tta aaa ccg cta ctt aat aca ctt ggg ctc      763
Ser Asp Tyr Lys Glu Asp Leu Lys Pro Leu Leu Asn Thr Leu Gly Leu
            170                 175                 180
ttt ttc caa att agg gat gat tat gct aat cta cac tcc aaa gaa tat      811
Phe Phe Gln Ile Arg Asp Asp Tyr Ala Asn Leu His Ser Lys Glu Tyr
        185                 190                 195
agt gaa aac aaa agt ttt tgt gaa gat ctg aca gag gga aag ttc tca      859
Ser Glu Asn Lys Ser Phe Cys Glu Asp Leu Thr Glu Gly Lys Phe Ser
    200                 205                 210
ttt cct act att cat gct att tgg tca agg cct gaa agc acc cag gtg      907
Phe Pro Thr Ile His Ala Ile Trp Ser Arg Pro Glu Ser Thr Gln Val
215                 220                 225                 230
cag aat atc ttg cgc cag aga aca gaa aac ata gat ata aaa aaa tac      955
Gln Asn Ile Leu Arg Gln Arg Thr Glu Asn Ile Asp Ile Lys Lys Tyr
                235                 240                 245
tgt gta cat tat ctt gag gat gta ggt tct ttt gaa tac act cgt aat     1003
Cys Val His Tyr Leu Glu Asp Val Gly Ser Phe Glu Tyr Thr Arg Asn
            250                 255                 260
acc ctt aaa gag ctt gaa gct aaa gcc tat aaa cag att gat gca cgt     1051
```

```
Thr Leu Lys Glu Leu Glu Ala Lys Ala Tyr Lys Gln Ile Asp Ala Arg
        265                 270                 275
ggt ggg aac cct gag cta gta gcc tta gta aaa cac tta agt aag atg     1099
Gly Gly Asn Pro Glu Leu Val Ala Leu Val Lys His Leu Ser Lys Met
    280                 285                 290
ttc aaa gaa gaa aat gaa taa tgttaagcca ttcttgattg gacctcatag       1150
Phe Lys Glu Glu Asn Glu  *
295             300
cttattttag ttaatctttn ntttgtcttt tagccttacc acctttttaaa aaatttgtta   1210
ttntccagaa acagtaaata ggtgagtagg ggtggtgcaa gtgaattcgt tttcatttag   1270
aagcccctct gtacagataa tcaaaattca aagttgaaag aatcaaaagc agccacagtt   1330
atgtaggtct gatttgaatg tcataattgc agtgacagga cattgccacc nnctcgtatc   1390
ctactaccat caatgttgtg tttattccgt caataaaaaa gacttgcttc caggaatttt   1450
tatccataca ctttctaact gtactatctg ggcagttcca agccagtttc tattagctag   1510
ctggaccaaa gaccacaaat ctctttttt cctaaac                             1547
```

<210> 4
<211> 300
<212> PRT
<213> Homo sapiens

<220>
<221> VARIANT
<222> 204
<223> diverging amino acid Leu in ref : GENESEQP R97565

<220>
<221> VARIANT
<222> 205
<223> diverging amino acid Gly in ref : GENESEQP R97565

<220>
<221> VARIANT
<222> 225
<223> diverging amino acid Ser in ref : GENESEQP R97565

<220>
<221> VARIANT
<222> 252
<223> diverging amino acid Lys in ref : GENESEQP R97565

<220>

<221> VARIANT
<222> 257
<223> diverging amino acid Gly in ref : GENESEQP R97565

<220>
<221> VARIANT
<222> 295
<223> diverging amino acid Ser in ref : GENESEQP R97565

<400> 4

```
Met Glu Lys Thr Gln Glu Thr Val Gln Arg Ile Leu Leu Glu Pro Tyr
1               5                  10                  15
Lys Tyr Leu Leu Gln Leu Pro Gly Lys Gln Val Arg Thr Lys Leu Ser
            20                  25                  30
Gln Ala Phe Asn His Trp Leu Lys Val Pro Glu Asp Lys Leu Gln Ile
        35                  40                  45
Ile Ile Glu Val Thr Glu Met Leu His Asn Ala Ser Leu Leu Ile Asp
    50                  55                  60
Asp Ile Glu Asp Asn Ser Lys Leu Arg Arg Gly Phe Pro Val Ala His
65                  70                  75                  80
Ser Ile Tyr Gly Ile Pro Ser Val Ile Asn Ser Ala Asn Tyr Val Tyr
                85                  90                  95
Phe Leu Gly Leu Glu Lys Val Leu Thr Leu Asp His Pro Asp Ala Val
            100                 105                 110
Lys Leu Phe Thr Arg Gln Leu Leu Glu Leu His Gln Gly Gln Gly Leu
        115                 120                 125
Asp Ile Tyr Trp Arg Asp Asn Tyr Thr Cys Pro Thr Glu Glu Glu Tyr
    130                 135                 140
Lys Ala Met Val Leu Gln Lys Thr Gly Gly Leu Phe Gly Leu Ala Val
145                 150                 155                 160
Gly Leu Met Gln Leu Phe Ser Asp Tyr Lys Glu Asp Leu Lys Pro Leu
                165                 170                 175
Leu Asn Thr Leu Gly Leu Phe Phe Gln Ile Arg Asp Asp Tyr Ala Asn
            180                 185                 190
Leu His Ser Lys Glu Tyr Ser Glu Asn Lys Ser Phe Cys Glu Asp Leu
        195                 200                 205
Thr Glu Gly Lys Phe Ser Phe Pro Thr Ile His Ala Ile Trp Ser Arg
    210                 215                 220
Pro Glu Ser Thr Gln Val Gln Asn Ile Leu Arg Gln Arg Thr Glu Asn
225                 230                 235                 240
Ile Asp Ile Lys Lys Tyr Cys Val His Tyr Leu Glu Asp Val Gly Ser
                245                 250                 255
```

Phe Glu Tyr Thr Arg Asn Thr Leu Lys Glu Leu Glu Ala Lys Ala Tyr
260 265 270
Lys Gln Ile Asp Ala Arg Gly Gly Asn Pro Glu Leu Val Ala Leu Val
275 280 285
Lys His Leu Ser Lys Met Phe Lys Glu Glu Asn Glu
290 295 300

<210> 5
<211> 49
<212> DNA
<213> Artificial sequence

<400> 5
aagtgaaatt ttcaattttt ttattagatt attattgaag tgacagaaa 49

<210> 6
<211> 50
<212> DNA
<213> Artificial sequence

<400> 6
aagtgaaatt ttcaattttt tttattagat tattattgaa gtgacagaaa 50

<210> 7
<211> 19
<212> DNA
<213> Artificial sequence

<400> 7
tgaaattttc aatttttttt 19

<210> 8
<211> 19
<212> DNA
<213> Artificial sequence

<400> 8
ctgagacttt cataatctg 19

<210> 9
<211> 20
<212> DNA

<213> Artificial sequence

<400> 9
atgagaccta ctgctaatcc 20

<210> 10
<211> 18
<212> DNA
<213> Artificial Sequence

<220>
<221> misc_binding
<222> 1..18
<223> sequencing oligonucleotide PrimerPU

<400> 10
tgtaaaacga cggccagt 18

<210> 11
<211> 18
<212> DNA
<213> Artificial Sequence

<220>
<221> misc_binding
<222> 1..18
<223> sequencing oligonucleotide PrimerRP

<400> 11
caggaaacag ctatgacc 18